# Characterizing the *Physcomitrella patens* ecotype Reute

Inaugural-Dissertation zur Erlangung der Doktorwürde

der Fakultät für Biologie

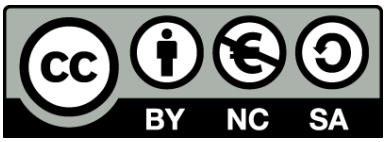
der Philipps-Universität Marburg


vorgelegt von

Manuel Hiß

Aus Zell am Harmersbach, Baden-Württemberg

Marburg, März 2019






Vom Fachbereich Biologie der Philipps-Universität Marburg (Hochschulkennziffer 1180) als Dissertation angenommen.

Erstgutachter: Prof. Dr. Stefan A. Rensing

Zweitgutachter: Prof. Dr. Alfred Batschauer

Dr. Katrin Heer

Prof. Dr. Annette Becker

Tag der Disputation am 25.06.2019.

*'Time is a drug. Too much of it kills you.'*

*Terry Pratchett*

# 1 Contents

# 2 Publications

## 2.1 Publications that contributed to this thesis

Essential results of this work were published as peer-reviewed articles. The publications are summarized in the chapter 6 “Results and Conclusions”, citing them as indicated below. Reprints of publications are included into this thesis on the indicated pages.

**Publication 1** **page 20**

Manuel Hiss, Oliver Laule, Rasa M Meskauskiene, Muhammad A Arif, Eva L Decker, Anika Erxleben, Wolfgang Frank, Sebastian T Hanke, Daniel Lang, Anja Martin, Christina Neu, Ralf Reski, Sandra Richardt, Mareike Schallenberg-Rüdinger, Peter Szövényi, Theodhor Tiko, Gertrud Wiedemann, Luise Wolf, Philip Zimmermann, Stefan A Rensing

Large-scale gene expression profiling data for the model moss *Physcomitrella patens* aid understanding of developmental progression, culture and stress conditions

The Plant Journal, 2014; 79 (3), 530-539

Own contribution: Conducted part of the microarray experiments, statistical and functional analysis of expression data, writing of parts of the manuscript. The work was supervised by S.A. Rensing.

**Publication 2** **page 31**

Manuel Hiss, Rabea Meyberg, Jens Westermann, Fabian B Haas, Lucas Schneider, Mareike Schallenberg-Rüdinger, Kristian K Ullrich, Stefan A Rensing

Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute

Plant Journal, 2017; 90(3):606-620

Own contribution: Conducted the microarray experiments, statistical and functional analysis of expression data, mapping of genomic DNA sequencing data, SNP data analysis, writing of parts of the manuscript. The work was supervised by K. K. Ullrich and S.A. Rensing.

Daniel Lang, Kristian K. Ullrich, Florent Murat, Jörg Fuchs, Jerry Jenkins, Fabian B. Haas, Mathieu Piednoel, Heidrun Gundlach, Michiel Van Bel, Rabea Meyberg, Cristina Vives, Jordi Morata, Aikaterini Symeonidi, Manuel Hiss, Wellington Muchero, Yasuko Kamisugi, Omar Saleh, Guillaume Blanc, Eva L. Decker, Nico van Gessel, Jane Grimwood, Richard D. Hayes, Sean W. Graham, Lee E. Gunter, Stuart McDaniel, Sebastian N.W. Hoernstein, Anders Larsson, Fay-Wei Li, Pierre-Francois Perroud, Jeremy Phillips, Priya Ranjan, Daniel S. Rokshar, Carl J. Rothfels, Lucas Schneider, Shengqiang Shu, Dennis W. Stevenson, Fritz Thümmler, Michael Tillich, Juan Carlos Villarreal A., Thomas Widiez, Gane Ka-Shu Wong, Ann Wymore, Yong Zhang, Andreas D. Zimmer, Ralph S. Quatrano, Klaus F.X. Mayer, David Goodstein, Josep M. Casacuberta, Klaas Vandepoele, Ralf Reski, Andrew C. Cuming, Jerry Tuskan, Florian Maumus, Jérome Salse, Jeremy Schmutz, Stefan A. Rensing

The *P. patens* chromosome-scale assembly reveals moss genome structure and evolution

The Plant Journal, 2018; 93 (3), 515-533

Own contribution: mapping of genomic DNA sequencing data, SNP data analysis, writing of parts of the manuscript. The work was supervised by K.K. Ullrich, P-F. Perroud and S.A. Rensing.



Manuel Hiss, Lucas Schneider, Christopher Grosche, Melanie A. Barth, Christina Neu, Aikaterini Symeonidi, Kristian K. Ullrich, Pierre-François Perroud, Mareike Schallenberg-Rüdinger and Stefan A. Rensing

Combination of the Endogenous *lhcsr1* Promoter and Codon Usage Optimization Boosts Protein Expression in the Moss *Physcomitrella patens*

Frontiers in Plant Science, 2017; Volume 8

Own contribution: Developed and tested the *in vivo* plate reader measurements, prepared parts of the transient constructs, performed parts of the transformation experiments, writing of the manuscript. The work was supervised by M. Schallenberg-Rüdinger, Kristian K. Ullrich, P-F. Perroud and S.A. Rensing.

**Additional publications not contributing to this thesis**

Kristian K Ullrich, Manuel Hiss, Stefan A Rensing
Means to optimize protein expression in transgenic plants
Current Opinion in Biotechnology, 2015; 32, 61-67
Own contribution: Writing part of the review manuscript

Mareike Schallenberg-Rüdinger, Bastian Oldenkott, Manuel Hiss, Phuong Le Trinh, Volker Knoop, Stefan A Rensing
A Single-Target Mitochondrial RNA Editing Factor of *Funaria hygrometrica* Can Fully Reconstitute RNA Editing at Two Sites in *Physcomitrella patens*
Plant and Cell Physiology, 2017; 58 (3), 496-507
Own contribution: Conducted the quantitative RT PCR and the statistical analysis of the expression data

Anja Possart, Tengfei Xu, Inyup Paik, Sebastian Hanke, Sarah Keim, Helen-Maria Hermann, Luise Wolf, Manuel Hiss, Claude Becker, Enamul Huq, Stefan A Rensing, Andreas Hiltbrunner
Characterization of Phytochrome Interacting Factors from the Moss *Physcomitrella patens* Illustrates Conservation of Phytochrome Signaling Modules in Land Plants
The Plant Cell, 2017; 29 (2), 310-330
Own contribution: Conducted part of the microarray experiments and the statistical and functional analysis of microarray expression data

Katrin Heer, Kristian K Ullrich, Manuel Hiss, Sascha Liepelt, Ralf Schulze Brüning, Jiabin Zhou, Lars Opgenoorth, Stefan A Rensing
Detection of somatic epigenetic variation in Norway spruce via targeted bisulfite sequencing
Ecology and Evolution, 2018, 8 (19), 9672–9682.
Own contribution: Prepared bisulphite sequencing libraries and contributed to the analysis and interpretation of the data

Pierre-François Perroud, Fabian B Haas, Manuel Hiss, Kristian K Ullrich, Alessandro Alboresi, Mojgan Amirebrahimi, Kerrie Barry, Roberto Bassi, Sandrine Bonhomme, Haodong Chen, Juliet Coates, Tomomichi Fujita, Anouchka Guyon-Debast, Daniel Lang, Junyan Lin, Anna Lipzen, Fabien Nogué, Melvin J Oliver, Inés Ponce de León, Ralph S Quatrano, Catherine Rameau, Bernd Reiss, Ralf Reski, Mariana Ricca, Younousse Saidi, Ning Sun, Peter Szövényi, Avinash Sreedasyam, Jane Grimwood, Gary Stacey, Jeremy Schmutz, Stefan A Rensing

The *Physcomitrella patens* gene atlas project: large scale RNA-seq based expression data

The Plant Journal 2018; 95(1), 168-182.

Own contribution: RNA extraction and contribution to analysis and interpretation of the data

## 2.2 Conference contributions

**Black Forest Summer School 2016**

Feldberg, Germany

Poster: Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute

**The 19th Annual Moss International Conference 2016**

Leeds, United Kingdom

Poster: Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: Introducing the ecotype Reute

**29th Conference on plant molecular biology 2015**

Dabringhausen, Germany

Poster: A closer look at ecotypes of the moss *Physcomitrella patens*

**The 18th Annual Moss International Conference 2015**

Cancún, México

Presentation: A closer look at the *Physcomitrella patens* 'Reute' strain

Poster: Measuring and optimizing expression strength in *Physcomitrella patens*

**The 17th Annual Moss International Conference 2014**

Chinese Normal University Beijing, China

Presentation: Large scale gene expression profiling data of the model moss *Physcomitrella patens* help to understand developmental progression, culture and stress conditions

Poster: Measuring and optimizing expression strength in *Physcomitrella patens*

Conference grant from the German Academic Exchange Service (DAAD)

**The 16th Annual Moss International Conference 2013**

Prague, Czech Republic

Poster: Influence of codon bias and promoter choice on expression strength in *Physcomitrella patens*

# Zusammenfassung

Für mehrere in der Forschung verwendete Pflanzenmodelle wurden ausgedehnte Ökotypsammlungen angelegt und diese verwendet um die genetischen Ursachen für phänotypische Unterschiede zu entschlüsseln und um die Anpassung und die Akklimatisierung der Ökotypen zu untersuchen. Auch für das Moos *Physcomitrella patens* sind Sammlungen aus geographisch getrennten Vorkommen vorhanden die für phylogenetische und taxonomische Studien verwendet wurden. Von diesen Pflanzen wurden aber nur wenige auf phänotypische Unterschiede untersucht oder genauer charakterisiert. Meine Doktorabeit beschäftigt sich mit *Physcomitrella patens* aus dem Fundort Reute bei Freiburg im Breisgau. Meine Arbeit zeigt, dass der am gängigsten verwendete Stamm aus Gransden Wood, Großbritannien signifikant weniger Sporophyten produziert als die Stämme aus Reute oder Villersexel, Frankreich obwohl sich die Gametangien bei allen drei Stämmen gleich schnell entwickeln und keine direkt sichtbaren phänotypischen Unterschiede aufweisen. Die Expressionsdaten aus verschiedenen Stadien der Sporophytenentwicklung im Stamm Reute und vergleichend dazu aus dem Stamm Gransden zeigen Unterschiede im Expressionsprofil mehrerer Gene und somit eine mögliche Ursache für die Unterschiede in der Sporophytenanzahl. Zur weiteren Charakterisierung der Stämme habe ich die genetischen Unterschiede mittels Genomsequenzierungen bestimmt und finde mehrere Genomabschnitte, die eine erhöhte Anzahl an Einzelnukleotidaustauschen aufweisen. Basierend auf meinen Ergebnissen empfehle ich *Physcomitrella patens* Reute als Ökotypen zu betrachten welcher sich besonders für Experimente zur Fortpflanzung und zum Generationenwechsel eignet.

Des Weiteren verwende ich Reute in molekularbiologischen Experimenten indem ich mittels Protoplastentransformation fluoreszente Proteine einbringe um den Effekt der Codonverwendung zu bestimmen und einen aktiven endogenen Promotor zu finden. Der in Pflanzen häufig verwendete 35S-Promotor zeigt in *P. patens* nur eine mittlere Aktivität und ist nicht konstitutiv exprimiert. Basierend auf Microarray Daten von Gransden und Reute habe ich den Promotor des LHCSR Gens ausgewählt, welches unter den meisten Versuchsbedingungen eine starke Expression aufweist. Um die Expressionsstärke zu bestimmen habe ich eine Fluoreszenz-Messmethode entwickelt, welche über einen Mikroplatten-Reader und ein zweites Fluoreszenzprotein zur Normalisierung der Messwerte, die Fluoreszenzintensität bestimmt. Der LHCSR Promotor zeigt in meinen Versuchen eine stärkere Expression als der zweifache 35S und der Actin Promotor und eine verkürzte Variante des Promotors zeigt ebenfalls die volle Promotoraktivität. Ich nutze ebenfalls den Effekt der Codonverwendung in *P. patens* auf die Expressionsstärke und stelle ein codonoptimiertes GFP zur Verfügung.

Zusammenfassend charakterisieren meine Ergebnisse den *P. patens* Ökotypen aus Reute und zeigen, dass er ein hilfreiches Werkzeug für reverse genetische Studien ist.

# 3 Abstract

Ecotype collections are used for several plant models to unravel the molecular causes of phenotypic differences and to investigate effects of environmental adaptation and acclimation. For the model moss *Physcomitrella patens*, collections of accessions are available and have been used e.g. for phylogenetic and taxonomic studies, but few were investigated further for phenotypic differences. My thesis focuses on the accession found in Reute close to Freiburg im Breisgau, Germany. My publication shows that the standard laboratory strain Gransden produces fewer sporophytes than Reute or Villersexel K3, although gametangia develop in the same time course and do not show evident morphological differences. My work provides expression profiling and comparative developmental data for several stages of sporophyte development, as well as information on genetic variation from genomic sequencing. There is variation in the expression profiles of several genes between Gransden and Reute, as well as genome segments that are variation hotspots. With my work I propose that Reute is considered a *P. patens* ecotype and suggest its use for investigations that involve progression through the life cycle and generational succession.

In my experiments I used the *P. patens* ecotype from Reute in molecular biology experiments introducing fluorescent proteins via chemical protoplast transformation to study codon usage bias and select a strong endogenous promoter. The 35S promoter, which is commonly used in plant systems, is only suitable to a limited extent in *Physcomitrella patens* due to mediocre and non-constitutive activity. Based on a broad range of Gransden and Reute microarray experiments we selected an LHCS gene that is highly expressed in most tissues and treatments. To measure expression strength I developed a novel fluorescence readout system, utilizing a microplate reader and an internal fluorescence control for normalization. The results from my publications demonstrate that the selected promoter is more active than the double 35S and Actin promoters. Deletion constructs were generated to develop a shorter promoter that retains the high activity of the full-length promoter. In parallel, codon bias within *P. patens* was analysed and demonstrated that the use of several codons is biased and correlates with expression strength. Two GFP variants were synthesized with different sets of codons and compared, showing that optimized codon usage increases the amount of protein under the control of strong promoters.

In summary the findings from my publications characterize the *P. patens* ecotype from Reute and demonstrate that it will be an useful tool in reverse genetic studies.

# 4 Introduction

## *4.1 Physcomitrella patens*

The moss *Physcomitrella patens* belongs to the family of *Funariaceae* and over the years has been developed as a model plant for evolutionary and developmental research (Jill Harrison, 2017, Rensing, 2017). *P. patens* has for some years been the only fully sequenced moss (Rensing *et al.*, 2008) thereby bridging the sequence information gap between the green algal model *Chlamydomonas reinhardtii* (Merchant *et al.*, 2007) and the seed plant model *Arabidopsis thaliana* (Arabidopsis Genome, 2000). It has fewer tissue types than seed plants and can be easily cultivated under axenic conditions (Beike *et al.*, 2010). The moss can regenerate into a whole plant from single cells, which is useful in the generation of transgenic lines via polyethylene glycol mediated protoplast transformation (Schaefer *et al.*, 1991, Cove *et al.*, 2009). In seed plants a targeted knock-out or insertion of genes has only recently been possible in a reliable way using the CRISPR/Cas system (Feng *et al.*, 2013, Li *et al.*, 2013, Nekrasov *et al.*, 2013, Shan *et al.*, 2013, Xie and Yang, 2013) which has also been shown to be efficient in *P. patens (Lopez-Obando et al., 2016)*. In contrast the high efficiency of homologous recombination in the multicellular eukaryote *P. patens* has allowed targeted genome modifications early in its research use (Schaefer and Zryd, 1997).

Figure 1 shows a schematic life cycle of *P. patens*. The dominant phase in the life cycle of the moss *P. patens* is the gametophytic phase. The haploid spore germinates and gives rise to protonema filaments. The protonema filaments consist of caulonema and chloronema cells. The chloronema cells contain many chloroplasts and show perpendicular cell walls whereas the caulonema cells show few chloroplasts and oblique cell walls. From the caulonema cells buds arise that will form into gametophores building a stem and phyllids. In cold conditions with low light intensities these gametophores will develop gametangia (Engel, 1968, Hohe *et al.*, 2002). The spermatozoids from the male antheridia will fertilize the female archegonia. The diploid sporophyte consisting of a seta and the spore capsule is formed and produces the haploid spores (Strotbek *et al.*, 2013). The development of the sporophyte can be divided into distinct stages which have been investigated in detail for the Gransden ecotype (Ortiz-Ramirez *et al.*, 2016).

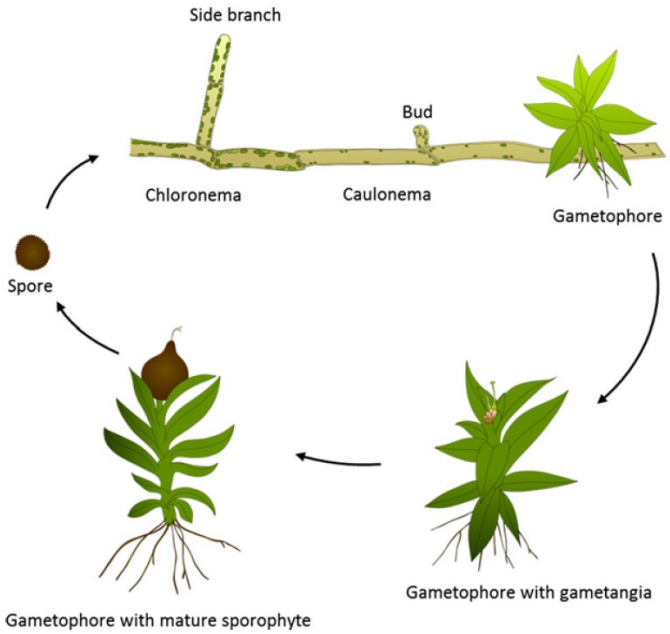



**Figure 1 The *Physcomitrella patens* life cycle.** Germination of haploid spores yields the juvenile gametophytic generation, the protonema. Protonema grows two-dimensional by apical (tip) growth and side branching. Protonemata consist of chloroplast-rich chloronema cells, and longer, thinner caulonema cells featuring less chloroplasts and oblique cross walls. Three-faced buds featuring single apical stem cells emerge from side branches (Harrison *et al.*, 2009) to form the adult gametophytic phase, the leafy gametophores. Gametophores comprise basal, multicellular rhizoids for nutrient supply, as well as non-vascular leaves (phyllids). Gametangia (female archegonia and male antheridia) develop on the gametophores. Upon fertilization of the egg cell by motile spermatozoids the diploid zygote forms and subsequently performs embryogenesis. Spore mother cells in the diploid sporophyte undergo meiosis to form spores.

## 4.2 Ecotypes

Ecotypes are genetically distinct geographical populations within a species, which are adapted to specific environmental conditions. In plant research ecotypes are studied e.g. to find and select traits that are beneficial for plant growth or harvest yields but also to elucidate the population distribution and progression of population development (Cao *et al.*, 2011). Several *P. patens* ecotypes have been described (von Stackelberg *et al.*, 2006) but only few have been sequenced or used for the generation of mutant lines. The commonly used strain was collected in Gransden Wood, UK (Engel, 1968) and the genomic sequence of this ecotype was published in 2008 (Rensing *et al.*, 2008). The ecotype collected near Villersexel, France (Villersexel K3) was used to generate a genetic linkage map (Kamisugi *et al.*, 2008) and stable mutant lines expressing GFP and mCherry (Gd-green and Vx-red; (Perroud *et al.*, 2011)). The ecotype from Kaskaskia Island, US has been sequenced in the frame of a cooperation between Monsanto and the Washington University in St. Louis. Gransden, Villersexel K3 and Kaskaskia can be crossed with each other, with the Gransden laboratory ecotype showing a low male fertility (Perroud *et al.*, 2011). The ecotype collected near Freiburg-Reute, Germany has been used for

generation of stable mutant lines expressing a GFP-tagged autophagy marker PpATG8 (Sanchez-Vera *et al.*, 2017).

## 4.3 Microarray analysis

DNA microarrays are a hybridisation-based method and enable genome-wide analysis of genetic variation and transcriptome-wide analysis of gene expression (Aharoni and Vorst, 2002). The technology compares e.g. gene expression strength between two or more samples. Different type of DNA microarrays are available based on the production method. Initially amplified cDNAs, expressed sequence tags (ESTs) or synthesised nucleic acids were spotted on paper as dot blots or reverse dot blots (Kafatos *et al.*, 1979, Saiki *et al.*, 1989) and later on microarray glass surfaces. Technical development and the availability of transcriptome sequencing information allowed the *in-situ* synthesis of DNA directly on the glass surfaces (Fodor *et al.*, 1991, Blanchard and Hood, 1996).

The microarray technology has been described in detail before (Aharoni and Vorst, 2002) and a short summary is given here. Each gene model is represented on the microarray by several short single strand sequences called probes that were selected to uniquely represent the gene model as probe set. Each short sequence is synthesised in one spot with a high copy number and the different spots of one probe set are distributed across the complete microarray. The RNA is extracted from the organism or tissue of interest and transcribed into cDNA. This cDNA is either directly amplified or transcribed into RNA and amplified into cRNA. The resulting cDNA or cRNA molecules are fragmented and labelled with a dye. These labelled fragments are hybridized to the microarray. Complementary sequences will bind to the probes on the microarray surface. Non-specifically bound sequences are washed of and a picture of the microarray surface is taken under light conditions that excite the used dye. The signal intensity of the microarray spots correlates with the number of transcripts in the sample and thereby with the expression strength of the corresponding gene.

This technology is used to compare the relative expression between different tissues or treatments since absolute expression levels can only be estimated up to a certain degree (Engelen *et al.*, 2006). The first spotted DNA microarray study in plant research focused on *A. thaliana* and compared mRNA samples from wild type and a transgenic line (Schena *et al.*, 1995). Commercially available microarrays cover several plant species like *A. thaliana*, *O. sativa*, *Z. mays* or *M. truncatula* (Galbraith, 2006). Till recently custom-made *P. patens* microarrays from Agilent, CombiMatrix and NimbleGen were commercially available which are based on sequence data from the Gransden ecotype. These arrays have been used in several studies (see *Table 1*) to investigate e.g. stress response (Cuming *et al.*, 2007), developmental stages (Wolf *et al.*, 2010, Busch *et al.*, 2013, O'Donoghue *et al.*, 2013, Hiss *et al.*, 2014, Beike *et al.*, 2015, Ortiz-Ramirez *et al.*, 2016) or expression patterns of different mutant lines (Komatsu *et al.*, 2013, Yaari *et al.*, 2015).

## 4.4 DNA sequencing and SNP analysis

The custom-made *P. patens* microarrays are no longer available and whole transcriptome gene expression analysis switched to using sequencing technology. Many current sequencing technologies focus on short DNA fragments and assemble them afterwards *in silico* to retrieve the complete sequence information. The different technologies have been summarized before (Metzker, 2010) and the Illumina sequencing technology is described here in more detail. With the Illumina sequencing technology the DNA is fragmented to sizes between 50 bp and 150 bp, adapters are added and the fragments are bound to complementary adapters on a glass surface. The fragments are amplified on the glass surface and form spots, with each spot containing only one sequence in a high copy number. After this preparation the DNA synthesis is started. A modified DNA polymerase incorporates one modified and labelled nucleotide into the complementary strand and is stopped by the blocked ends of the modified nucleotide. A picture is taken after each nucleotide incorporation and based on the colour of the spot the nucleotide identity can be determined. The modified nucleotide is chemically unblocked and the next nucleotide can be incorporated. After image analysis and sequence determination the short sequences are either assembled into longer fragments or mapped to an existing genome or transcriptome sequence. To detect single nucleotide variations (SNPs) for each position in a genome, the number of unique short reads covering the position are compared to the reference genome. Based on these counts the genotype can be estimated (Heather and Chain, 2016, Shendure *et al.*, 2017). The Illumina technology has been developed by Solexa and has been introduced by resequencing the human genome (Bentley *et al.*, 2008). For *P. patens* research mainly Illumina sequencing has been used (see Table *1*) which is also the case for the sequencing done on *P. patens* as part of the Gene Atlas project (http://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/) and as part of this thesis.

**Table 1 List of published microarray and RNA-seq datasets for *Physcomitrella patens***. Study topic and type of analysis summarize the publication goals; third column lists the source material for RNA extraction; technology column names the used microarray or sequencing technology; fifth column lists the gene annotation version used.

| study topic | type of analysis | RNA extracted from | technology | annotation version | publication |
|---|---|---|---|---|---|
| ABA treatment and water stress | pertubation | protonema | Agilent microarray | 3' biased ends Physcobase | Cuming et al., 2007 |
| dessication tolerance | pertubation | protonema | Agilent microarray | JGI v1.1 | Komatsu et al., 2013 |
| sporophyte development | development | protonema, green sporophyte | Agilent microarray | JGI v1.1 | O'Donoghue |
| UV light | pertubation | gametophores | Combimatrix microarray | cosmoss v1.2 | Wolf et al., 2010 |
| phyllid development | development | detached phyllids | Combimatrix microarray | cosmoss v1.2 | Busch et al., 2013 |
| developmental stages, biotic stress, light intensity and quality, water stress, culture media | pertubation, development | protoplasts, spores, protonema, gametophores | Combimatrix microarray | cosmoss v1.2 | Hiss et al., 2015 |
| cold treatment | pertubation | gametophores | Combimatrix microarray | cosmoss v1.2 | Beike et al., 2015 |
| transcriptome atlas | development | protoplasts, spores, chloronema, caulonema, gametophores, rhizoids, archegonia, | Nimblegen microarray | cosmoss v1.6 | Ortiz-Ramirez et al., 2015 |
| DNA methylation KO mutant | mutant analysis | protonema | Nimblegen microarray | cosmoss v1.6 | Yaari et al., 2015 |
| heat dependent alternative splicing | pertubation | protonema | Illumina Hiseq 2000 | cosmoss v1.6 | Chang et al., 2014 |
| light dependent alternative splicing | pertubation | protonema | Illumina Hiseq 2000 | cosmoss v1.6 | Wu et al., 2014 |
| DEK1 KO mutant | mutant analysis | protonema / gametophores | Illumina Hiseq 2000 | cosmoss v1.6 | Demko et al., 2014 |
| NAC (VNS7) overexpression mutant | mutant analysis | protonema | Illumina GAIIx | cosmoss v1.6 | Xu et al., 2014 |
| phyllid development | development | detached phyllids | AB SOLiD System 2.0 | JGI v1.1 | Nishiyama et al., 2012 |
| phyllid development WOX13 double mutant | development | detached phyllids | AB SOLiD System 2.0 | JGI v1.1 | Sakakibara et al., 2014 |
| red light response of PUBS HY2 double mutant | mutant analysis | protonema | Illumina GAIIx | cosmoss v1.6 | Chen et al., 2012 |
| gametophore development | development | gametophores, chloronema, caulonema | Illumina GAIIx | cosmoss v1.6 | Xiao et al., 2011 |
| protonema regeneration | development | protoplasts | Illumina GAIIx | cosmoss v1.6 | Xiao et al., 2012 |
| single cell transcriptomes during shoot developm | development | protonema tip cells, gametophore buds, whole plants | Illumina Hiseq 2000 | cosmoss v1.6 | Frank et al., 2015 |
| abiotic stress | pertubation | protonema | Illumina Hiseq 2000 | cosmoss v1.6 | Khraiwesh et al., 2015 |

## 4.5 Transformation

In eukaryotic cells transformation can be used to introduce sequences of interest into the target organism. Targeted insertion into the genome via homologous recombination is only efficient in few eukaryotes and was first discovered in yeast (Hinnen *et al.*, 1978). In *P. patens* a high efficiency for targeted insertion by homologous recombination was found (Schaefer *et al.*, 1991). Several methods to introduce DNA fragments into eukaryotic cells have been developed and in *P. patens* the protoplast transformation with polyethylene glycol and calcium chloride is the most common used method (Schaefer *et al.*, 1991). To generate protoplasts the plant cell wall is digested by fungal enzymes. The protoplasts are centrifuged, the enzyme mix is washed away and the cells are kept in an osmotically adjusted sugar solution. In the polyethylene glycol transformation protocol the polyethylene glycol then perforates the cell membrane of the protoplasts to allow the introduction of the DNA fragments. After the transformation, the polyethylene glycol is washed off and the protoplasts can be put on a regenerative medium containing glucose to allow them to re-build their cell wall. If an antibiotic resistance gene was part of the introduced DNA fragment, the regenerated protoplasts are put on selective medium that contains antibiotics to identify positive transformants. Since transient transformants may survive the selective medium the survivors of the first round of selection are put on medium without selection pressure. The transiently transformed DNA fragments will be removed from the cytoplasm as soon as the plant can survive without them. Afterwards the plants are put on selective medium again allowing the selection of lines with a stable integration into the genomic sequence (Egener *et al.*, 2002, Cove *et al.*, 2009, Liu and Vidali, 2011).

## 4.6 Fluorescent reporter proteins

Fluorescent reporter genes are used in different species to localize proteins, show protein interactions or measure promoter strength. Since the origin of many reporter genes lies within aquatic animals the reporter genes were mutated stepwise in the lab to enable their use in non-aquatic species. The gene for the commonly used green fluorescent protein was found and isolated from jellyfish (Chalfie *et al.*, 1994, Inouye and Tsuji, 1994) and has been mutagenized to select variants with higher fluorescence intensities or different excitation and emission wavelength (Heim *et al.*, 1995, Tsien, 2005). This was reached by small structural changes to the chromophore-adjacent amino acids but in the case of fluorescence intensity also by improving solubility induced by changes in amino acids facing the outer surface of the protein as summarized in(Zacharias and Tsien, 2006). For plants a soluble modified GFP showed high intensities (Davis and Vierstra, 1998). This GFP variant is used also in *P. patens* were it shows higher fluorescence intensities than the modified GFP (Cho *et al.*, 1999).

## 4.7 Codon usage

The amino acid sequence of proteins is determined by blocks of three nucleotides (codons) during protein synthesis. Since there are more codons than necessary to code for the common amino acids some of the amino acids are encoded by several codons (Crick *et al.*, 1961). The usage of these redundant codons is not the same between different organisms (Grantham, 1978). This may be caused or is reflected in the copy number of the respective tRNA genes (Higgs and Ran, 2008). As a consequence of this codon usage bias the expression strength of a gene may differ between organisms even if the same nucleotide sequence is expressed. In some cases the organism may even lack the necessary tRNA for certain codons and will not be able to express genes that use that specific codon. Therefore a codon optimization of gene sequences can increase expression strength if a gene sequence from one organism is introduced into another for research or biotechnological purposes (Itakura *et al.*, 1977). This codon bias can also be observed within the gene set of one organism and is thought as a mechanism to regulate gene expression (Sharp *et al.*, 1993). There are several reports of a positive correlation between expression strength and usage of certain codons in *Caenorhabditis elegans*, *Drosophila melanogaster*, and *A. thaliana* (Duret and Mouchiroud, 1999). Codon usage has been calculated for *P. patens* based on EST data (Stenoien, 2005) and based on the published genome sequence (Rensing *et al.*, 2008, Zimmer *et al.*, 2013, Lang *et al.*, 2018). The codon usage bias in *P. patens* seems to be driven by a combination of weak natural selection and the predominant mutational biases (Szovenyi *et al.*, 2017).

# 5 Research objectives

This thesis aims to introduce the *Physcomitrella patens* ecotype Reute as a useful addition to the commonly used ecotype Gransden and the ecotype Villersexel K3. In my first publication I prepared and analysed whole transcriptome microarray data for several developmental stages and perturbations on the ecotype Gransden and use the Gransden microarray as well as for samples from both ecotypes Reute and Villersexel K3 (Publication 1). The Reute ecotype has been initially described but has rarely been used for standard molecular biology methods. In my second publication I introduced the ecotype in more detail. This included a description of both morphological, phenological and molecular differences between Gransden and Reute (Publication 2). We set up sporophyte experiments to confirm and analyse differences in sporophyte development between Gransden and Reute that several laboratories observed within the last years. To include gene expression data into the analysis of Gransden and Reute differences I prepared whole transcriptome microarrays for several Reute tissues with a focus on sporophyte development. Finally I confirmed the genetic difference

between the ecotypes by including an analysis of the single nucleotide polymorphisms between the three ecotypes Gransden, Reute and Villersexel K3. In a recent publication with my contribution (Publication 3) I added the analysis of a fourth ecotype from Kaskaskia Island and we put the SNP analyses from all ecotypes into the context of the chromosome scale assembly of the *P. patens* Gransden genome.

In my third publication I used the *P. patens* ecotype Reute to test if expression data generated for Gransden can be used to support experiments in Reute and if they allow prediction of expression strength of genes across ecotypes (Publication 4). This was done by developing and testing a plate reader measurement system for promoter-reporter gene constructs in a transient *in vivo* transformation background. This measurement system was used to test a strong endogenous promoter for the use in *P. patens* and investigate if a modified GFP variant that is codon optimized for the *P. patens* codon bias can provide stronger GFP signals. Our experiments also tested if the Reute ecotype can be used for molecular biology experiments and can be as easily transformed as Gransden.

# 6 Results and Conclusions

## 6.1 Publication 1

My first publication shows that the microarray platform based on the genomic sequence of the *P. patens* ecotype Gransden can be used to measure expression profiles of the Reute and Villersexel K3 ecotypes of which I included one triplicate each. This publication also provides extensive microarray expression data on a broad range of developmental stages as well as different perturbations for the Gransden ecotype. This created the basis for comparisons to the other ecotypes Reute and Villersexel K3 but also allows comparisons to other species for which expression profiling data are available.

the plant journal

SEB Society for Experimental Biology

*The Plant Journal* (2014) **79,** 530–539 doi: 10.1111/tpj.12572

RESOURCE

# Large-scale gene expression profiling data for the model moss *Physcomitrella patens* aid understanding of developmental progression, culture and stress conditions

**Manuel Hiss[1,2,3], Oliver Laule[4,†], Rasa M. Meskauskiene[4], Muhammad A. Arif[5,6], Eva L. Decker[5], Anika Erxleben[5,‡], Wolfgang Frank[6], Sebastian T. Hanke[1,2], Daniel Lang[5], Anja Martin[5,§], Christina Neu[2], Ralf Reski[3,5,7,8], Sandra Richardt[2,¶], Mareike Schallenberg-Rüdinger[1], Peter Szövényi[2,**], Theodhor Tiko[2], Gertrud Wiedemann[5], Luise Wolf[2,**,††], Philip Zimmermann[4,*] and Stefan A. Rensing[1,2,3,7,*]**

[1]*Plant Cell Biology, Faculty of Biology, University of Marburg, Karl-von-Frisch-Strasse 8, 35043 Marburg, Germany,*
[2]*Faculty of Biology, University of Freiburg, Schänzlestrasse 1, 79104 Freiburg, Germany,*
[3]*FRISYS Freiburg Initiative for Systems Biology, University of Freiburg, 79104 Freiburg, Germany,*
[4]*NEBION AG, Hohlstrasse 515, 8048 Zurich, Switzerland,*
[5]*Department of Plant Biotechnology, Faculty of Biology, University of Freiburg, Schänzlestrasse 1, 79104 Freiburg, Germany,*
[6]*Plant Molecular Cell Biology, Department Biology I, Ludwig-Maximilians-University Munich, LMU Biocenter, Grosshadernerstrasse 2-4, 82152 Planegg-Martinsried, Germany,*
[7]*BIOSS Centre for Biological Signalling Studies, University of Freiburg, 79104 Freiburg, Germany and*
[8]*FRIAS – Freiburg Institute for Advanced Studies, University of Freiburg, 79104 Freiburg, Germany*

Received 1 August 2013; revised 22 May 2014; accepted 27 May 2014; published online 6 June 2014.
*For correspondence (e-mails stefan.rensing@biologie.uni-marburg.de or phz@nebion.com).
‡Present address: Department of Pharmaceutical Bioinformatics, Faculty of Chemistry, Pharmacy, and Earth Sciences, University of Freiburg, Hermann-Herder-Strasse 9, 79104 Freiburg, Germany.
§Present address: Thermo Fisher, Opelstrasse 9, 68789 St Leon-Rot, Germany.
¶Present address: Roche Diagnostics GmbH, Nonnenwald 2, 82377 Penzberg, Germany.
**Present address: Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.
††Present address: Department of Biozentrum, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland.
†deceased.

## SUMMARY

**The moss *Physcomitrella patens* is an important model organism for studying plant evolution, development, physiology and biotechnology. Here we have generated microarray gene expression data covering the principal developmental stages, culture forms and some environmental/stress conditions. Example analyses of developmental stages and growth conditions as well as abiotic stress treatments demonstrate that (i) growth stage is dominant over culture conditions, (ii) liquid culture is not stressful for the plant, (iii) low pH might aid protoplastation by reduced expression of cell wall structure genes, (iv) largely the same gene pool mediates response to dehydration and rehydration, and (v) AP2/EREBP transcription factors play important roles in stress response reactions. With regard to the AP2 gene family, phylogenetic analysis and comparison with *Arabidopsis thaliana* shows commonalities as well as uniquely expressed family members under drought, light perturbations and protoplastation. Gene expression profiles for *P. patens* are available for the scientific community via the easy-to-use tool at https://www.genevestigator.com. By providing large-scale expression profiles, the usability of this model organism is further enhanced, for example by enabling selection of control genes for quantitative real-time PCR. Now, gene expression levels across a broad range of conditions can be accessed online for *P. patens*.**

**Keywords: *Physcomitrella patens*, moss, gene expression, microarray, genevestigator, transcriptomics, development, culture, stress.**

## INTRODUCTION

Gene expression profiles are a valuable community resource. They allow researchers interested in certain sets of genes or conditions (tissues, developmental stages, stress treatments, etc.) to investigate transcription levels *in silico* and to generate hypotheses to subsequently put to the test. Through the availability of expression data in compliance with the minimum information about a microarray experiment (MIAME) standard (Brazma *et al.*, 2001; Zimmermann *et al.*, 2006) in repositories such as ARRAYEXPRESS (Rocca-Serra *et al.*, 2003), incorporation into meta-analysis tools such as GENEVESTIGATOR (Hruz *et al.*, 2008) becomes feasible. The availability of data in such a tool allows for the end user to browse with ease through experiments conducted by different labs or using different technology platforms. Moreover, the use of anatomy, development and treatment ontologies allows users to analyze, for example, developmental progression or to compare expression data across taxonomic boundaries.

The moss *Physcomitrella patens* has been developed over the last decade as a plant model organism for which a large set of experimental tools is available (Reski and Cove, 2004; Frank *et al.*, 2005; Quatrano *et al.*, 2007; Kamisugi *et al.*, 2008; Lang *et al.*, 2008; Prigge and Bezanilla, 2010; Mueller *et al.*, 2014). The sequencing of the genome (Rensing *et al.*, 2008) allowed the development of a gene expression microarray covering all predicted protein-coding genes (Wolf *et al.*, 2010).

We decided to generate expression profiles from a set of principal tissues/developmental stages and environmental/stress treatments that we consider useful for the community. Here we present the generation of large-scale gene expression data for *P. patens* as well as their integration and availability via GENEVESTIGATOR.

## RESULTS

The initial set of expression profile data represents a wide range of conditions, including various tissue types, stages of development and perturbations (Table 1). In general, three biological replicates were generated per condition. Exceptions are the leaflet time series where each time point is represented by a single biological replicate (Busch *et al.*, 2013) and the developmental stage 'gametophore formed' with four biological replicates. To verify the level of standardization of experimental conditions we analyzed all samples using hierarchical clustering. The results show a close clustering of biological replicates relative to other samples, confirming a high level of reproducibility (Figure S1). In the following we have conducted some example analyses to demonstrate the usefulness of the data to analyze development, culture conditions and stress.

**Table 1** Experimental conditions. List of experimental conditions that are available in GENEVESTIGATOR as microarray data sets. The ARRAYEXPRESS bulk accession numbers for the experiments are given. Each experiment consists of three biological replicates with the exception of the time series of detached leaflets (one replicate for each time point) and the developmental stage 'gametophore formed' (four replicates)

| | Tissue/treatment | Bulk accession number |
|---|---|---|
| Tissues/cells (Anatomy) | Spores | E-MTAB-916 |
| | Protoplasts | E-MTAB-976 |
| | Protonemata | E-MTAB-976, E-MTAB-917 |
| | Gametophores | E-MTAB-917, E-MTAB-916 |
| Developmental stages (Development) | Germination (protonemal development) | |
| | Germinated spores | E-MTAB-916 |
| | Primarily chloronemata | E-MTAB-976, E-MTAB-917 |
| | Gametophore growth | |
| | Gametophore formed | E-MTAB-917 |
| | Gametangia development | |
| | Mature antheridia/archegonia | E-MTAB-917 |
| | Different genotype (Reute, Villersexel) | E-MTAB-916 |
| Perturbations | | |
| Biotic | Botrytiscinerea inoculation | E-MTAB-919 |
| Light intensity | Darkness | E-MTAB-913 |
| | Strong light | E-MTAB-913 |
| | Sunlight | E-MTAB-913 |
| Light quality | UV-B (303 nm) supplementation, different intensities | E-MEXP-2508 |
| | Shift long day to short day | E-MTAB-917 |
| Water stress | Dehydration | E-MTAB-914 |
| | Rehydration | E-MTAB-914 |
| Other | Different growth media | E-MTAB-976 |
| | Shift from pH 5.8 to pH 4.5 | E-MTAB-976 |
| | Timeseries of detached leaflets | E-MTAB-915 |

### Growth stage is dominant over culture form

Principal component analysis across all experiments from the initial set of published data shows a deep cleft between liquid culture/filamentous (protonemal) growth stage and culture on solid medium/gametophore stage. There are 229 genes differentially expressed (Cyber-T, $q < 0.05$) between protonema on liquid versus solid medium, while 247 genes are differentially expressed between protonema and gametophores on solid medium. Partial least squares analysis confirms that the majority of differences in expression are covariant with the two principal growth stages, protonema and gametophores (Figure 1), which are therefore dominant over culture form with regard to alteration of the gene expression profile.

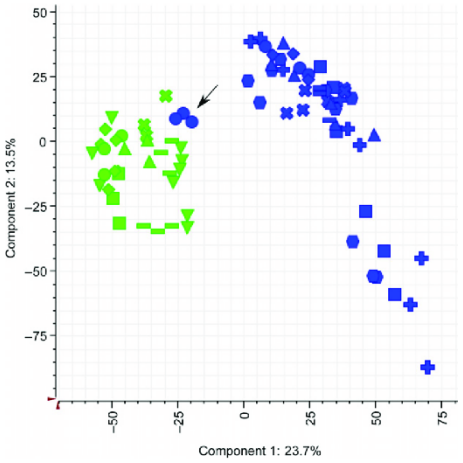



**Figure 1.** Partial least squares analysis across experiments.
The tissue type is used as the covariant (activity) factor. The plot shows in blue the experiments on a solid medium (mainly gametophores) and in green the liquid culture experiments (protonemata). Arrow shows the three replicate experiments for protonema on solid medium. Growth stage (tissue type) is clearly dominant over culture form, since these replicates cluster with the other protonema experiments, although those were conducted in liquid culture. Axes show the first two components and the covariance explained by these components.

### Liquid culture does not represent a stress condition

Although many genes are differentially regulated between protonema in liquid culture and on solid medium, stress-related Gene Ontology categories are not over-represented among these. Therefore, the regular shearing (fragmentation) of protonemal filaments done during the liquid culture regime does not seem to constitute a stress condition, unlike, for example, darkness or dehydration (see below), where stress terms are found to be over-represented among the differentially regulated genes.

### An acidic medium might aid protoplastation by downregulation of cell wall structure genes

It is recommended that plant material which will be used for protoplast preparation is grown at pH 4.5 prior to enzymatic digestion of the cell wall (Hohe and Reski, 2002). Plants growing at pH 4.5 apparently have a different cell wall composition from plants growing at pH 5.8 (the cell wall can be more easily digested to produce protoplasts, as measured by a higher rate of released protoplasts). Comparing the gene expression data for tissue growing at pH 5.8 and at pH 4.5 should give insights in the activation of genes responsible for the change in cell wall composition.

Six hundred and seventy-one genes are found to be upregulated by the shift to pH 4.5, and 270 downregulated (Cyber-T, $q < 0.05$). Among the upregulated genes there are only three that are annotated as involved in 'cell wall catabolism', but no expansins (Schipper *et al.* 2002) or other genes associated with cell wall loosening, like xyloglucan endotransglucosylases or pectinesterases (Lagaert *et al.*, 2009), are found. In contrast, the downregulated genes include five expansin genes and three cellulose synthase genes, resulting in an overrepresentation of the Gene Ontology (GO) term 'plant-type cell wall organization' (Table 2). This observation seems to clash with previous studies in cucumber (*Cucumis sativus*), where expansin proteins were shown to increase activity after the external pH was changed to acidic conditions (McQueen-Mason *et al.*, 1992). However, reaction to acidic conditions might vary between different plant species, and the control treatment was growth at pH 5.8, i.e. mildly acidic conditions. In the case of *P. patens* the downregulation of specific cell wall structure genes might influence cell wall composition, leading to better digestibility of the cell wall. Another explanation might be that only newly formed cells start to change their cell wall composition at pH 4.5 (Hohe and Reski, 2002).

### Dehydration and rehydration responses are mediated by the same gene pool

Statistical analysis (see Experimental Procedures) of the gene expression data for all expressed genes from dehydrated (50% fresh weight loss), rehydrated and untreated gametophores finds 690 genes upregulated and 1231 genes downregulated after an hour of dehydration out of a total of 26 853 genes on the microarray. Genes important in regulation of transcription, protein modification and

**Table 2** Gene Ontology (GO) term enrichment analysis – pH shift.
List of significantly overrepresented GO terms (biological process ontology) and corresponding *q*-values from the enrichment analysis for genes downregulated after a pH shift from 5.8 to 4.5

| GO-term | Number of genes | *q*-Value |
|---|---|---|
| Unsaturated fatty acid biosynthetic process | 5 | $1.7 \times 10^{-4}$ |
| Glucose metabolic process | 9 | $5.0 \times 10^{-3}$ |
| Plant-type cell wall organization | 4 | $9.1 \times 10^{-3}$ |
| Fatty acid biosynthetic process | 5 | $2.3 \times 10^{-2}$ |
| Spermidine biosynthetic process | 2 | $2.5 \times 10^{-2}$ |
| Carbon fixation | 3 | $2.5 \times 10^{-2}$ |
| NAD biosynthetic process | 2 | $3.1 \times 10^{-2}$ |
| Pyridine nucleotide biosynthetic process | 2 | $3.8 \times 10^{-2}$ |
| Glucose catabolic process | 6 | $4.1 \times 10^{-2}$ |
| Lipid biosynthetic process | 7 | $4.1 \times 10^{-2}$ |
| Fatty acid metabolic process | 5 | $4.4 \times 10^{-2}$ |
| Polyamine metabolic process | 2 | $4.6 \times 10^{-2}$ |

**Table 3** Gene Ontology (GO) term enrichment analysis – dehydration and rehydration. List of significantly overrepresented GO terms (biological process ontology) and corresponding *q*-values from the enrichment analysis for genes upregulated after dehydration

| GO term | Number of genes | *q*-Value |
|---|---|---|
| Regulation of transcription, DNA-dependent | 37 | $4.2 \times 10^{-4}$ |
| Regulation of transcription | 38 | $4.2 \times 10^{-4}$ |
| Protein modification process | 57 | $3.7 \times 10^{-3}$ |
| Protein ubiquitination | 18 | $3.7 \times 10^{-3}$ |
| Metal ion transport | 12 | $1.3 \times 10^{-2}$ |
| Response to water | 2 | $2.0 \times 10^{-2}$ |
| Protein amino acid dephosphorylation | 5 | $3.2 \times 10^{-2}$ |
| SRP-dependent cotranslational protein targeting to membrane | 2 | $3.2 \times 10^{-2}$ |
| Lipid metabolic process | 14 | $3.2 \times 10^{-2}$ |
| Cation transport | 16 | $3.2 \times 10^{-2}$ |
| Mitochondrial transport | 3 | $3.2 \times 10^{-2}$ |
| Exocytosis | 3 | $3.2 \times 10^{-2}$ |
| Cell adhesion | 3 | $3.2 \times 10^{-2}$ |
| Oligosaccharide metabolic process | 3 | $3.2 \times 10^{-2}$ |
| Guanosine tetraphosphate metabolic process | 2 | $3.2 \times 10^{-2}$ |
| Lipid glycosylation | 2 | $3.2 \times 10^{-2}$ |
| Ion transport | 17 | $3.8 \times 10^{-2}$ |
| Steroid metabolic process | 2 | $4.0 \times 10^{-2}$ |

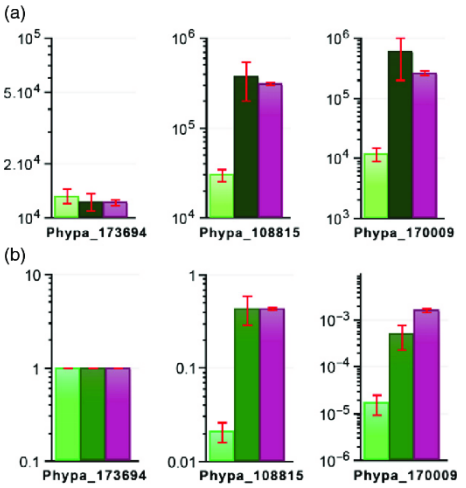



**Figure 2.** Bar charts of normalized expression strength.
Normalized expression level of the *lea* genes Phypa_108815 and Phypa_170009 and of the reference gene Phypa_173694 (encoding a thioredoxin) according to microarray (a) and quantitative real time PCR (b). Light green, control sample; dark green, dehydration sample; lilac, rehydration sample. Error bars show the standard error of the mean of two to four biological replicates. The *y*-axis in (a) is in arbitrary fluorescence units scaled to a median of 10 000, and in (b) arbitrary fluorescence units normalized to the reference gene Phypa_173694.

response to water are significantly overrepresented among the upregulated genes after dehydration ($q$-value < 0.05; Table 3). As previously shown for protonema (Cuming *et al.*, 2007), effector genes like *lea* (late embryogenesis abundant; also known as dehydrins), were found to be upregulated after dehydration. An association of *lea* proteins with osmotic stress and response to ABA has also been described for mosses and seed plants in different studies (Kamisugi and Cuming, 2005; Olvera-Carrillo *et al.*, 2010). The two *lea* genes Phypa_108815 and Phypa_170009 are strongly expressed after dehydration and remain expressed after 1 h of rehydration (Figure 2a). They were used in this study to validate the microarray gene expression by quantitative real-time PCR (Figure 2b). As shown before for this platform (Busch *et al.* 2013) the data are in very good agreement.

The large-scale expression data enabled us to find reference genes on a more global scale than previously possible. We selected a new reference gene (Phypa_173694, a thioredoxin gene; Figure 2), which shows stable expression over all conducted microarray expression analyses (Figure S2). In contrast, out of 12 reference genes used in previous studies, and selected for phytohormone treatments in a recently published study (Le Bail *et al.*, 2013), only ARC34 (Phypa_146870) shows a globally stable expression over all microarray experiments (Figure S2). Due to the diverse set of conditions the present data are therefore well suited for the selection of reference genes allowing us to investigate a wide range of stages and perturbations.

Besides the two *lea* genes, 619 of the 690 genes upregulated after dehydration remain activated after rehydration. The same key stress regulator and effector genes activated during dehydration and rehydration might be explained by the comparable stress situations during those two treatments that represent the changes in the water regime that poikilohydric plants have to cope with. Activation during dehydration and rehydration was also seen for some *lea* genes in *Tortula* (*Syntrichia*) *ruralis* (Oliver *et al.*, 2005). The 71 genes which are upregulated after dehydration and downregulated after rehydration are linked to 'lipid biosynthetic processes' and acting in 'cytoskeleton organization' and 'lipid metabolic processes' as shown in GO term enrichment analyses ($q$-value < 0.05). Genes within these groups mainly encode for membrane repair proteins. This leads to the suggestion that the main damage to the membranes has already been repaired during dehydration, whereas most processes, like the downregulation of photosynthesis-related genes, are active during dehydration and

rehydration (Table S1). Our dehydration and rehydration data on gametophores extend the existing data on dehydration of protonemata (Cuming *et al.*, 2007) by adding rehydration and by detecting more genes as being differentially expressed under drought conditions.

### Phylogenetic and comparative analysis of stress-mediating AP2/EREBP transcription factors

Of the 690 genes upregulated after dehydration, 126 were predicted to be responsive to ABA (Timmerhaus *et al.*, 2011), including members of the AP2/EREBP transcription factor family (Lang *et al.*, 2010). The AP2/EREBP family is involved in both salt stress and ABA responses in *P. patens* (Richardt *et al.*, 2010). Members of the AP2/EREBP family are also detected by ANOVA as upregulated (Figure 3, Table S2) in all of the available non-standard light conditions (strong light, sunlight and UV light). This evidence strengthens the suggestion that AP2/EREBP factors have a central regulatory role during stress conditions.

To detect evolutionarily conserved expression patterns in the AP2/EREBP family the differentially expressed genes (DEGs) were annotated in a phylogenetic tree of the gene family based on members from *Arabidopsis thaliana* (167 sequences), *Chlamydomonas reinhardtii* (14 sequences) and *P. patens* (156 sequences). Of 156 *Physcomitrella* sequences, 39 were detected by ANOVA with post-hoc testing as DEGs under the tested conditions (sunlight, strong light, darkness, UV-B, drought, pH shift, protonema in liquid culture and protoplastation). Interestingly, we noticed that one subclade within the tree contains most (69%) of the *P. patens* AP2-DEGs. These DEGs show diverse expression patterns under the different tested conditions (Figure 4). Only a few Arabidopsis sequences (10% of all sequences) and no *Chlamydomonas* sequences are found within this subclade. This particular subclade shows *P. patens* genes that are activated under several stress conditions, like Pp1s373_18V6.1 and Pp1s60_269V6.1 which are activated under protoplastation as well as drought, UV-B and sunlight (Figure 4). Such genes potentially represent upstream mediators that integrate different stress response pathways. In that regard they are similar to *A. thaliana* genes from the same subclade of the AP2 family which are activated under many or all of these stresses, most prominent among them being At3g50260.1 [encoding COOPERATIVELY REGULATED BY ETHYLENE AND JASMONATE 1 (CEJ1), also known as DREB AND EAR MOTIF PROTEIN 1 (DEAR 1)], which has been described as being involved in several stress pathways, namely cold, drought and defense to bacteria (Lamesch *et al.*, 2012). Based on its presence in the same subclade, and having a similar activation profile, we suggest that Pp1s60_269V6.1 might play a similar role in *P. patens*.

Within this subclade there are also more specialized genes that are induced only under specific conditions, like Pp1s60_228V6.1 after UV-B exposure or Pp1s199_50V6.1 during protoplastation (Figure 4). Again in the lower part of the tree there are also *A. thaliana* genes that show such a specific profile, for example At5g67190.1 (*DEAR2*), a close paralog of *DEAR1*, that is transcriptionally activated in protoplasts (Figure 4). Two *P. patens* genes, Pp1s199_50V6.1 and Pp1s240_84V6.1, show similar activation. Such genes are candidates for regulators that act downstream and thus mediate more specific responses, in this case stresses involved in protoplastation.

### Integration into and availability in the GENEVESTIGATOR tool

In order to integrate the *P. patens* data into GENEVESTIGATOR (https://www.genevestigator.com), application ontologies were developed for tissue types, developmental stages and experimental factors (see Experimental Procedures) by adapting standard ontologies developed in collaboration with Plant Ontology (Walls *et al.*, 2012; Cooper *et al.*, 2013). GENEVESTIGATOR is not a microarray data analysis tool per se, but is a gene expression search engine that focuses on integrating the complete content and comparing results between experiments. As a starting point *P. patens* gene IDs (look-up between release versions and cross-links to other databases are provided online) can be selected, or sets of experiments. As an illustration, we used the Perturbations tool from the Gene Search toolset to identify the top five genes that are most strongly upregulated in individual perturbations but show minimal regulation in all other conditions. We then clustered these genes according to their expression profile in the perturbation and

| | Strong light | | | | Climate chamber | | | | UV light | | | | Dark | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TAP | | AP2/EREBP | | TAP | | AP2/EREBP | | TAP | | AP2/EREBP | | TAP | | AP2/EREBP | |
| | up | down | up | down | up | down | up | down | up | down | up | down | up | down | up | down |
| Sun light | 13 | 7 | 4 | 0 | 53 | 5 | 8 | 1 | 111 | 48 | 18 | 11 | 162 | 11 | 17 | 1 |
| Strong light | | | | | 55 | 1 | 7 | 0 | 116 | 45 | 18 | 9 | 161 | 8 | 19 | 1 |
| Climate chamber | | | | | | | | | 54 | 94 | 11 | 15 | 139 | 11 | 10 | 2 |
| UV light | | | | | | | | | | | | | 118 | 47 | 15 | 14 |

Low High

**Figure 3.** Number of up- and downregulated transcriptional regulator genes and AP2/EREBP family members.
As found by ANOVA using the factor light intensity. Red colors show a low number of genes and green colors a high number of genes. Transcription associated proteins (TAPs) comprise transcription factors and general transcriptional regulators (Lang *et al.*, 2010). Genes are shown as up- and downregulated compared with the treatment in the first column (e.g. in strong light there are 13 TAPs upregulated and seven TAPs downregulated compared with in sunlight).

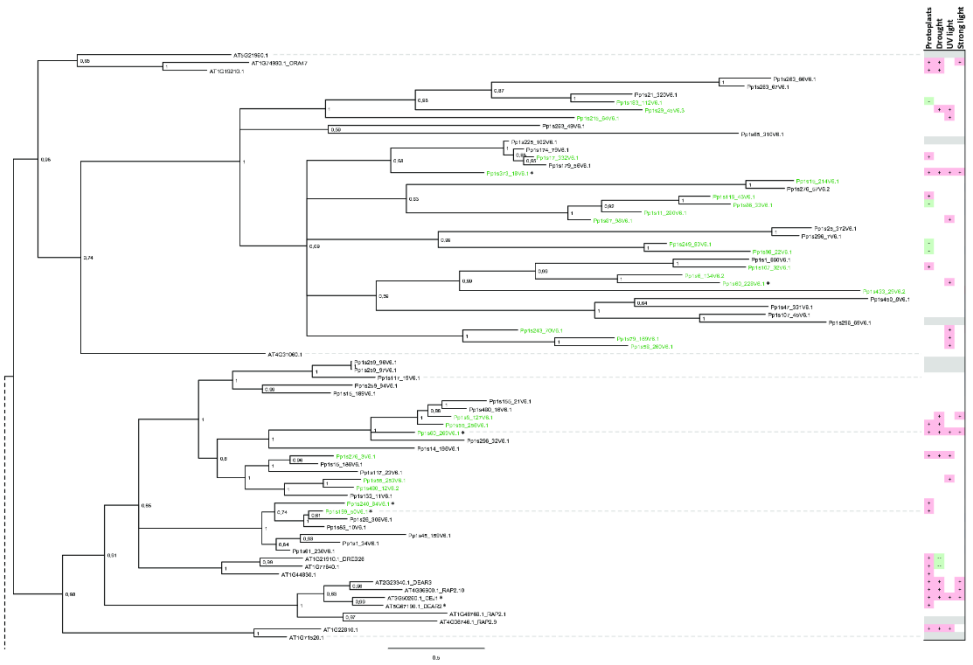



**Figure 4.** Rooted phylogenetic tree of AP2/EREBP family proteins.
Phylogenetic tree of part of the *Arabidopsis thaliana* (15 sequences) and *Physcomitrella patens* (60 sequences) AP2/EREBP family members; numbers at the nodes are support values (posterior probabilities from MRBAYES). The heatmap shows if the genes were detected via microarray analysis as upregulated (+/red) or downregulated (−/green) in the corresponding condition (see Experimental Procedures and Table S4 for *A. thaliana*) as compared with the control (see Experimental Procedures). No expression data were available for genes with gray bars. Green-colored identifiers are genes that were detected as differentially expressed in the above or one of the following additional conditions: sunlight, darkness, protonema in liquid culture (supplemented medium) and protonema in liquid culture at pH 4.5. Genes that are marked with an asterisk are mentioned in the Results: Pp1s373_18V6.1, Pp1s60_629V6.1, Pp1s60_228V6.1, Pp1s199_50V6.1, Pp1s240_84V6.1, At3g50260.1 (*DEAR1*/*CEJ1*) and At5g67190.1 (*DEAR2*).

development matrices using the Hierarchical Clustering tool. The results show a clustering of conditions that share gene-specificity profiles (Figure 5a), i.e. genes that were specifically upregulated in the chosen conditions but are unchanged in all other conditions. The clusters represented in Figure 5 are responsive to sunlight, protoplastation, photoperiod, UV-B, biotic stress and dehydration/rehydration, respectively. If these genes are plotted against tissue types/developmental stages (Figure 5b), their clustering reveals several distinct groups of genes that have quite different expression domains. The Gene Search tools in GENEVESTIGATOR further allow the identification of genes that have properties as defined by the user, for example being specifically expressed in a tissue, at a particular stage of development or in response to a perturbation. The search is performed by comparing the average expression in a target category (e.g. a chosen tissue type) with the sum of average expressions from a baseline set of categories (e.g. all tissue types). This approach allows us to look for genes that are generally specific for a chosen category (i.e. as compared with all other categories) or relatively specific (as compared with only a subset of categories). Due to the nature of the underlying data, such comparisons can only be performed between categories of the same type, such as tissues against tissues or perturbations against perturbations. Intuitive interfaces with checkboxes to choose categories of interest make it very straightforward for users to run this type of analysis.

## DISCUSSION

*Physcomitrella patens* has been established as an important model for plant evolutionary developmental biology (e.g. Tanahashi *et al.*, 2005; Menand *et al.*, 2007; Mosquna *et al.*, 2009; Okano *et al.*, 2009; Khandelwal *et al.*, 2010; Khraiwesh *et al.*, 2010; Sakakibara *et al.*, 2013) and comparative genomics (e.g. O'Toole *et al.*, 2008; Rensing *et al.*, 2008; Peers *et al.*, 2009; Cutler *et al.*, 2010; Perez-Rodriguez *et al.*, 2010; Richardt *et al.*, 2010). In collaboration between

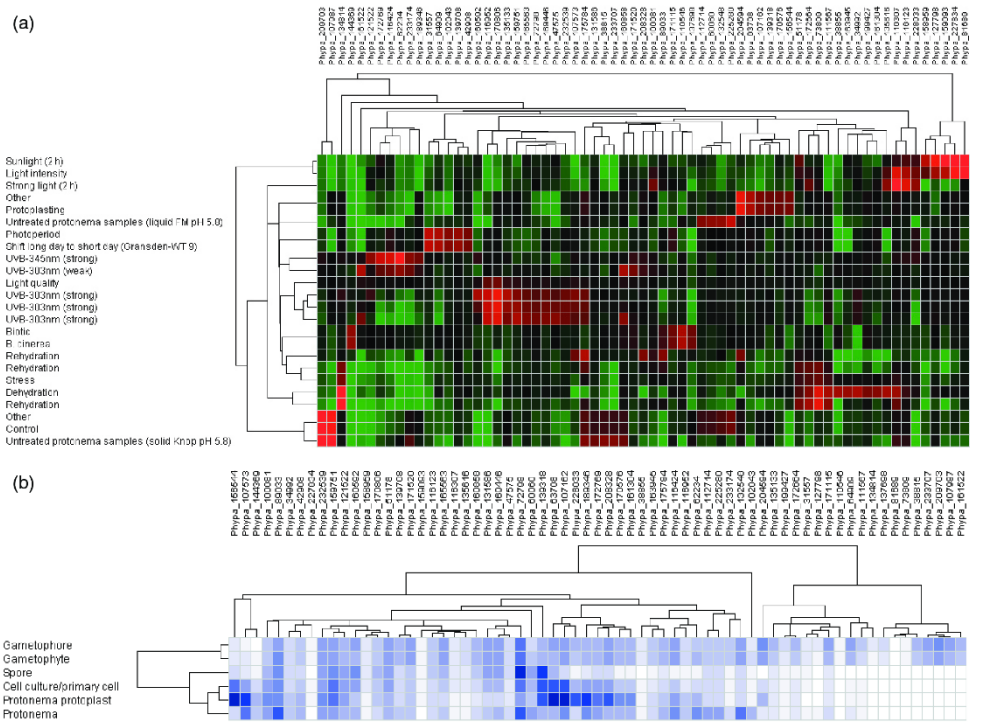



**Figure 5**. Snapshots of *Physcomitrella patens* data clustered in GENEVESTIGATOR.
(a) The Gene Search Perturbations tool was used to identify genes that are specifically upregulated in individual experimental conditions. The expression matrix from the combined list of most specific genes was clustered by genes and by perturbation type. Upregulation is shown in green, downregulation in red.
(b) The same list of genes was clustered by absolute expression across different stages of *P. patens* development. The more intense the blue color, the higher the gene expression.

the US Department of Energy and the International *Physcomitrella* Genome Consortium, the V3 plant flagship genome representation is a work in progress (pre-release available at http://phytozome.org/). Given the large and ever-increasing interest in this plant model, it was high time to create and make available expression profiling data for *P. patens*. We are confident that *P. patens* is a valuable addition to GENEVESTIGATOR that will not only represent a resource for research on *P. patens* as such but will also enhance cross-species comparisons of gene expression among photosynthetically active species. As shown, for example, for the dehydration and rehydration and culture condition analyses in this study, genes can be identified with specific functions under selected conditions. Also, expression profiles of transcription factors and their phylogenetic comparison with other plants (as shown for the example of AP2 here) can be used to find candidate regulators that coordinate responses to several stimuli or regulate specific pathways. The data can also be used to identify promoters that act under discrete conditions or to derive global control genes, for example for quantitative real-time PCR. The currently released data are soon to be complemented with more experimental conditions from the presented *P. patens* microarray platform (manuscripts in preparation). Moreover, a new publicly available microarray design has been generated based on the improved V1.6 gene annotation (Zimmer *et al.*, 2013) and *P. patens* is part of an RNA-seq pilot study conducted with the US Department of Energy to derive further expression profiles of development, stress and metabolic perturbations.

## EXPERIMENTAL PROCEDURES

### Plant material and growth conditions

The *P. patens* laboratory strain 'Gransden 2004' (Rensing *et al.*, 2008) was used for the majority of experiments. Since this strain

was subjected to selfing approximately once a year, an experiment using a parental strain, six selfing cycles away, was also conducted. In addition, experiments were carried out using the *P. patens* isolates 'Reute' and 'Villersexel' (McDaniel *et al.*, 2010), both displaying more sexual vigor than the Gransden strain. The former is genetically very close to Gransden (McDaniel *et al.*, 2010) while the latter exhibits a significant amount of polymorphism (von Stackelberg *et al.*, 2006) and has been used as the second parental strain for genetic mapping (Kamisugi *et al.*, 2008).

Plants were grown under long-day conditions (16 h white light, 8 h dark) in Knop medium at 20–25°C as previously described (Wolf *et al.*, 2010). Exceptions are listed in Table 1, summarizing all the xperimental conditions. Isolation of RNA and microarray processing were carried out as previously described, including array platform and design information (Wolf *et al.*, 2010).

### Mapping gene IDs

The *P. patens* microarray expression data are based on V1.2 gene models. Phylogenetic analyses were conducted with the V1.6 gene models. Conversion between the identifiers was done using the mapping information available on http://cosmoss.org/.

### Statistical testing

Microarray data processing was carried out as previously described (Wolf *et al.*, 2010). To detect differentially expressed genes the Cyber-T test was performed (Long *et al.*, 2001). All false discovery rate (FDR) corrections were carried out as described by Benjamini and Hochberg (1995). One-way analysis of variance (one-way ANOVA) used the hydration state or light intensity as factor and effect, respectively. States were defined as high (rehydration), low (dehydration) or normal (untreated control). For light intensities photosynthetically active radiation (UV-B) was used (see Table S3). An ANOVA post-hoc test was used to correct for multiple testing and calculated the FDR-corrected *P*-values (*q*-values) for all possible state combinations. The GO term enrichment analyses used Fisher's exact test to calculate *P*-values. Multiple testing corrected (Benjamini and Hochberg, 1995) *q*-values were calculated in R with the function p.adjust (R Development Core Team, 2008). Partial least squares (PLS) analysis used culture condition and tissue type as potential responses and analyzed their covariance with the activity factor tissue type. ANOVA with a post-hoc test, Cyber-T, hierarchical clustering and PLS were carried out with ANALYST 7.5 (Genedata, https://www.genedata.com/). The GO term enrichment analyses were conducted using in-house scripts. Clustering and visualization was carried out using GENEVESTIGATOR or ANALYST.

### Phylogenetic analysis

The selection of sequences was done using an existing nucleic acid sequence-based phylogeny of all genes detected to be AP2 transcription factors (from *A. thaliana*, *C. reinhardtii* and *P. patens*) based on classification rules previously described (Lang *et al.*, 2010). Genes detected as differentially expressed in *P. patens* were annotated in the tree and the subclade containing most of the *P. patens* DEGs was selected for further analyses.

For this subtree (which did not contain *C. reinhardtii* sequences), the corresponding *P. patens* V1.6 protein sequences were retrieved from http://cosmoss.org/ and the *A. thaliana* sequences from TAIR 10 and aligned with MAFFT-LINSI (v.7.037b, http://mafft.cbrc.jp/alignment/software/). The alignment was manually curated with JALVIEW (v.2.8, http://www.jalview.org/). PROTTEST (v.3.3, http://code.google.com/p/prottest3/) was used to select the most suitable substitution model (JTT+I+G+F). The phylogenetic tree was constructed with the MRBAYES (v.3.2.2 × 64, http://mrbayes.sourceforge.net/) parallelized version using the above-mentioned model with eight gamma distributed rates, two hot and two cold chains and 50 burn-in trees. The run was stopped after the standard deviation of split frequencies dropped below 0.01 (1.4 million generations and with no remaining observable trend detectable in the overlay plot). The protein sequence subtree was rooted based on the outgroup information from the nucleotide tree. The curated alignment is available upon request.

Expression data and fold-change matrices for *A. thaliana* and *P. patens* were retrieved from GENEVESTIGATOR. For *A. thaliana* several studies existed for each of the conditions, and if one or more experiments showed an up- or downregulation it was marked with a + or −, respectively, in the heatmap visualization (Figure 4). Control experiments for *P. patens* fold changes were protonemata at pH 5.8 for the protoplasts and gametophores in the developmental stage 'gametophore formed' for drought, UV light and strong light.

### Quantitative real-time PCR

For quantitative real-time PCR, RNA was reverse transcribed using SuperScript III (Invitrogen, http://www.invitrogen.com) and random hexamer primers (Fermentas, http://www.thermoscientific-bio.com/fermentas/). PRIMER3 (Untergasser *et al.*, 2012) was used for the design of specific oligonucleotides. Primer sequences used for amplification of the respective gene models are available upon request. For each 20-µl reaction, 20 ng of reverse-transcribed RNA was used and the reaction was carried out using SensiMix dT and SYBRGreen (Invitrogen) on a PicoReal Real-Time-PCR System (Thermo Scientific, http://www.thermoscientific.com). The concentration of cDNA was normalized to a thioredoxin transcript (Phypa_173694), showing expression-level and treatment-independent expression over all microarray analyses (cf. Experimental Procedures; Figure S2). The thioredoxin transcript was selected using a coefficient of variance filtering of the normalized mean expression values. Triplicate measurements were performed for each of two to three biological replicates. Analyses were performed with EXPRESSIONIST ANALYST 7.5 (Genedata).

### Development of application ontologies for GENEVESTIGATOR integration

Data available in GENEVESTIGATOR are manually curated using a controlled vocabulary from sample description ontologies. In order to integrate the *P. patens* experimental data into GENEVESTIGATOR, application ontologies were developed by adapting standard ontologies developed in collaboration with the Plant Ontology Consortium (POC, released May 2011 on http://www.plantontology.org/; http://wiki.plantontology.org/index.php/Summary_of_Changes_to_PO_May_2011) and with the aid of expert knowledge. The ontologies, in particular the perturbation ontologies, are highly dynamic and will be adapted according to the growing database content. The current ontologies comprise 57 anatomical categories, 18 developmental stages and 33 defined perturbation-related comparisons. Many of those are already available as experimental data (Table 1).

### Availability

All data have been made available in the public repository ARRAYEXPRESS (http://www.ebi.ac.uk/arrayexpress/) under the accession numbers shown in Table 1. The array design described here has recently been replaced by a new design (Nimblegen_Ppat_

SR_exp_HX12: Nimblegen 12 × 135 k chip, four 60mer probes per gene, V1.6 gene models) that is publicly available. The GENEVESTIGATOR tool and supporting documentation is available at https://www.genevestigator.com/gv/.

## ACKNOWLEDGEMENTS

Funding by the German Research Foundation (RE 837/10-2 to RR and SAR) and by the German Federal Ministry of Education and Research (FKZ 0315057B to AM and GW and Freiburg Initiative for Systems Biology, FKZ 0313921, to RR and SAR) is gratefully acknowledged. Financial support of the DAAD to PS and MAA, and of the Alexander von Humboldt Foundation and the SNSF to PS, is also acknowledged. We are grateful to Stefanie Pilz for excellent technical assistance.

## CONFLICT OF INTEREST

OL, RMM and PZ are/were employed by Nebion, the company providing GENEVESTIGATOR.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

**Table S1.** Photosynthesis related genes (according to Gene Ontology annotation) downregulated during dehydration and rehydration.

**Table S2.** Members of the AP2/EREBP transcription factor family up- and downregulated by different light treatments.

**Table S3.** Experimental conditions of sun light, strong light, UV light, dehydration and rehydration and darkness.

**Table S4.** *Arabidopsis thaliana* studies chosen from GENEVESTIGATOR.

**Figure S1.** Hierarchical clustering of microarray experiments.

**Figure S2.** Bar charts of microarray expression values from published reference genes and Phypa_173694 (thioredoxin) in various conditions.

## REFERENCES

**Benjamini, Y. and Hochberg, Y.** (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* **57**, 289–300.

**Brazma, A., Hingamp, P., Quackenbush, J.** *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371.

**Busch, H., Boerries, M., Bao, J., Hanke, S.T., Hiss, M., Tiko, T. and Rensing, S.A.** (2013) Network theory inspired analysis of time-resolved expression data reveals key players guiding P. patens stem cell development. *PLoS One*, **8**, e60494.

**Cooper, L., Walls, R.L., Elser, J.** *et al.* (2013) The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol.* **54**, e1.

**Cuming, A.C., Cho, S.H., Kamisugi, Y., Graham, H. and Quatrano, R.S.** (2007) Microarray analysis of transcriptional responses to abscisic acid and osmotic, salt, and drought stress in the moss, *Physcomitrella patens*. *New Phytol.* **176**, 275–287.

**Cutler, S.R., Rodriguez, P.L., Finkelstein, R.R. and Abrams, S.R.** (2010) Abscisic acid: emergence of a core signaling network. *Annu. Rev. Plant Biol.* **61**, 651–679.

**Frank, W., Decker, E.L. and Reski, R.** (2005) Molecular tools to study *Physcomitrella patens*. *Plant Biol. (Stuttg)*, **7**, 220–227.

**Hohe, A. and Reski, R.** (2002) Optimisation of a bioreactor culture of the moss *Physcomitrella patens* for mass production of protoplasts. *Plant Sci.* **163**, 69–74.

**Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W. and Zimmermann, P.** (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinformatics*, **2008**, 420747.

**Kamisugi, Y. and Cuming, A.C.** (2005) The evolution of the abscisic acid-response in land plants: comparative analysis of group 1 LEA gene expression in moss and cereals. *Plant Mol. Biol.* **59**, 723–737.

**Kamisugi, Y., von Stackelberg, M., Lang, D., Care, M., Reski, R., Rensing, S.A. and Cuming, A.C.** (2008) A sequence-anchored genetic linkage map for the moss, *Physcomitrella patens*. *Plant J.* **56**, 855–866.

**Khandelwal, A., Cho, S.H., Marella, H., Sakata, Y., Perroud, P.F., Pan, A. and Quatrano, R.S.** (2010) Role of ABA and ABI3 in desiccation tolerance. *Science*, **327**, 546.

**Khraiwesh, B., Arif, M.A., Seumel, G.I., Ossowski, S., Weigel, D., Reski, R. and Frank, W.** (2010) Transcriptional control of gene expression by microRNAs. *Cell*, **140**, 111–122.

**Lagaert, S., Belien, T. and Volckaert, G.** (2009) Plant cell walls: protecting the barrier from degradation by microbial enzymes. *Semin. Cell Dev. Biol.* **20**, 1064–1073.

**Lamesch, P., Berardini, T.Z., Li, D.** *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210.

**Lang, D., Zimmer, A.D., Rensing, S.A. and Reski, R.** (2008) Exploring plant biodiversity: the *Physcomitrella* genome and beyond. *Trends Plant Sci.* **13**, 542–549.

**Lang, D., Weiche, B., Timmerhaus, G., Richardt, S., Riano-Pachon, D.M., Correa, L.G., Reski, R., Mueller-Roeber, B. and Rensing, S.A.** (2010) Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.* **2**, 488–503.

**Le Bail, A., Scholz, S. and Kost, B.** (2013) Evaluation of reference genes for RT qPCR analyses of structure-specific and hormone regulated gene expression in *Physcomitrella patens* gametophytes. *PLoS One*, **8**, e70998.

**Long, A.D., Mangalam, H.J., Chan, B.Y., Tolleri, L., Hatfield, G.W. and Baldi, P.** (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. *J. Biol. Chem.* **276**, 19937–19944.

**McDaniel, S.F., von Stackelberg, M., Richardt, S., Quatrano, R.S., Reski, R. and Rensing, S.A.** (2010) The speciation history of the Physcomitrium-Physcomitrella species complex. *Evolution*, **64**, 217–231.

**McQueen-Mason, S., Durachko, D.M. and Cosgrove, D.J.** (1992) Two endogenous proteins that induce cell wall extension in plants. *Plant Cell*, **4**, 1425–1433.

**Menand, B., Yi, K., Jouannic, S., Hoffmann, L., Ryan, E., Linstead, P., Schaefer, D.G. and Dolan, L.** (2007) An ancient mechanism controls the development of cells with a rooting function in land plants. *Science*, **316**, 1477–1480.

**Mosquna, A., Katz, A., Decker, E.L., Rensing, S.A., Reski, R. and Ohad, N.** (2009) Regulation of stem cell maintenance by the Polycomb protein FIE has been conserved during land plant evolution. *Development*, **136**, 2433–2444.

**Mueller, S.J., Lang, D., Hoernstein, S.N.** *et al.* (2014) Quantitative analysis of the mitochondrial and plastid proteomes of the moss *Physcomitrella patens* reveals protein macrocompartmentation and microcompartmentation. *Plant Physiol.* **164**, 2081–2095.

**Okano, Y., Aono, N., Hiwatashi, Y., Murata, T., Nishiyama, T., Ishikawa, T., Kubo, M. and Hasebe, M.** (2009) A polycomb repressive complex 2 gene regulates apogamy and gives evolutionary insights into early land plant evolution. *Proc. Natl Acad. Sci. USA*, **106**, 16321–16326.

**Oliver, M.J., Velten, J. and Mishler, B.D.** (2005) Desiccation tolerance in bryophytes: a reflection of the primitive strategy for plant survival in dehydrating habitats? *Integr. Comp. Biol.* **45**, 788–799.

**Olvera-Carrillo, Y., Campos, F., Reyes, J.L., Garciarrubio, A. and Covarrubias, A.A.** (2010) Functional analysis of the group 4 late embryogenesis abundant proteins reveals their relevance in the adaptive response during water deficit in *Arabidopsis*. *Plant Physiol.* **154**, 373–390.

**O'Toole, N., Hattori, M., Andres, C., Iida, K., Lurin, C., Schmitz-Linneweber, C., Sugita, M. and Small, I.** (2008) On the expansion of the pentatricopeptide repeat gene family in plants. *Mol. Biol. Evol.* **25**, 1120–1128.

**Peers, G., Truong, T.B., Ostendorf, E., Busch, A., Elrad, D., Grossman, A.R., Hippler, M. and Niyogi, K.K.** (2009) An ancient light-harvesting protein is critical for the regulation of algal photosynthesis. *Nature*, **462**, 518–521.

**Perez-Rodriguez, P., Riano-Pachon, D.M., Correa, L.G., Rensing, S.A., Kersten, B. and Mueller-Roeber, B.** (2010) PlnTFDB: updated content and

new features of the plant transcription factor database. *Nucleic Acids Res.* **38**, D822–D827.

**Prigge, M.J. and Bezanilla, M.** (2010) Evolutionary crossroads in developmental biology: *Physcomitrella patens*. *Development*, **137**, 3535–3543.

**Quatrano, R.S., McDaniel, S.F., Khandelwal, A., Perroud, P.F. and Cove, D.J.** (2007) *Physcomitrella patens*: mosses enter the genomic age. *Curr. Opin. Plant Biol.* **10**, 182–189.

**R Development Core Team** (2008) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

**Rensing, S.A., Lang, D., Zimmer, A.D. *et al.*** (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.

**Reski, R. and Cove, D.J.** (2004) Quick guide: *Physcomitrella patens*. *Curr. Biol.* **14**, R261–R262.

**Richardt, S., Timmerhaus, G., Lang, D., Qudeimat, E., Correa, L.G., Reski, R., Rensing, S.A. and Frank, W.** (2010) Microarray analysis of the moss *Physcomitrella patens* reveals evolutionarily conserved transcriptional regulation of salt stress and abscisic acid signalling. *Plant Mol. Biol.* **72**, 27–45.

**Rocca-Serra, P., Brazma, A., Parkinson, H. *et al.*** (2003) ArrayExpress: a public database of gene expression data at EBI. *C R Biol.* **326**, 1075–1078.

**Sakakibara, K., Ando, S., Yip, H.K., Tamada, Y., Hiwatashi, Y., Murata, T., Deguchi, H., Hasebe, M. and Bowman, J.L.** (2013) KNOX2 genes regulate the haploid-to-diploid morphological transition in land plants. *Science*, **339**, 1067–1070.

**Schipper, O., Schaefer, D., Reski, R. and Flemin, A.** (2002) Expansins in the bryophyte Physcomitrella patens. *Plant molecular biology*, **50**, 789–802.

**von Stackelberg, M., Rensing, S.A. and Reski, R.** (2006) Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. *BMC Plant Biol.* **6**, 9.

**Tanahashi, T., Sumikawa, N., Kato, M. and Hasebe, M.** (2005) Diversification of gene function: homologs of the floral regulator FLO/LFY control the first zygotic cell division in the moss *Physcomitrella patens*. *Development*, **132**, 1727–1736.

**Timmerhaus, G., Hanke, S.T., Buchta, K. and Rensing, S.A.** (2011) Prediction and validation of promoters involved in the abscisic acid response in *Physcomitrella patens*. *Mol. Plant*, **11**, 11.

**Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G.** (2012) Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115.

**Walls, R.L., Athreya, B., Cooper, L. *et al.*** (2012) Ontologies as integrative tools for plant science. *Am. J. Bot.* **99**, 1263–1275.

**Wolf, L., Rizzini, L., Stracke, R., Ulm, R. and Rensing, S.A.** (2010) The molecular and physiological response of *Physcomitrella patens* to UV-B radiation. *Plant Physiol.* **153**, 1123–1134.

**Zimmer, A.D., Lang, D., Buchta, K., Rombauts, S., Nishiyama, T., Hasebe, M., Van de Peer, Y., Rensing, S.A. and Reski, R.** (2013) Reannotation and extended community resources for the genome of the non-seed plant *Physcomitrella patens* provide insights into the evolution of plant gene structures and functions. *BMC Genomics*, **14**, 498.

**Zimmermann, P., Schildknecht, B., Craigon, D. *et al.*** (2006) MIAME/Plant – adding value to plant microarrray experiments. *Plant Methods*, **2**, 1.

## 6.2 Publication 2

With my second publication I show that Reute and Villersexel K3 produce higher numbers of sporophytes if compared to Gransden but otherwise do not show phenotypic differences. The comparison of the microarray expression profiles between Gransden and Reute provide evidence that differences are also visible on the gene expression level. This included several transcription associated proteins that are differentially expressed between Gransden and Reute. The microarray expression data on sporophyte development of the Reute ecotype also show differentially regulated transcription factors and transcription regulators and additionally genes coding for cell wall modifying proteins like pectin methylesterase and xylosyltransferase. The expression profiling on sporophytes has recently been extended using sequencing technology (Perroud *et al.,* 2018). With this publication I also provide genomic sequencing data for the Reute ecotype and compare the SNPs found between the ecotypes. We find for Villersexel K3 a SNP density comparable to SNP densities found between *Arabidopsis thaliana* ecotypes with the Reute ecotype showing a much lower SNP density and therefore being genetically closer to Gransden.

the plant journal

SEB Society for Experimental Biology

*The Plant Journal* (2017) **90**, 606–620 doi: 10.1111/tpj.13501

RESOURCE

# Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute

**Manuel Hiss[1], Rabea Meyberg[1], Jens Westermann[1,†], Fabian B. Haas[1], Lucas Schneider[1,‡], Mareike Schallenberg-Rüdinger[1,§], Kristian K. Ullrich[1,¶] and Stefan A. Rensing[1,2,*]**

[1]*Plant Cell Biology, Faculty of Biology, University of Marburg, Karl-von-Frisch-Str. 8, 35043, Marburg, Germany,*
[2]*BIOSS Centre for Biological Signaling Studies, University of Freiburg, Freiburg, Germany, and*

Received 1 August 2016; revised 20 January 2017; accepted 24 January 2017; published online 5 February 2017.
*For correspondence (e-mail stefan.rensing@biologie.uni-marburg.de).
†Present address: Biocenter, Botanical Institute, University of Cologne, Zülpicherstr. 47b, 50674, Cologne, Germany.
‡Present address: Institute for Transfusion Medicine and Immunohematology, Johann-Wolfgang-Goethe University and German Red Cross Blood Service, Sandhofstraße 1, 60528, Frankfurt am Main, Germany.
§Present address: IZMB–Institut für Zelluläre und Molekulare Botanik, Abteilung Molekulare Evolution, Universität Bonn, Kirschallee 1, 53115, Bonn, Germany.
¶Present address: Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306, Ploen, Germany.

## SUMMARY

**Rich ecotype collections are used for several plant models to unravel the molecular causes of phenotypic differences, and to investigate the effects of environmental adaption and acclimation. For the model moss *Physcomitrella patens* collections of accessions are available, and have been used for phylogenetic and taxonomic studies, for example, but few have been investigated further for phenotypic differences. Here, we focus on the Reute accession and provide expression profiling and comparative developmental data for several stages of sporophyte development, as well as information on genetic variation via genomic sequencing. We analysed cross-technology and cross-laboratory data to define a confident set of 15 mature sporophyte-specific genes. We find that the standard laboratory strain Gransden produces fewer sporophytes than Reute or Villersexel, although gametangia develop with the same time course and do not show evident morphological differences. Reute exhibits less genetic variation relative to Gransden than Villersexel, yet we found variation between Gransden and Reute in the expression profiles of several genes, as well as variation hot spots and genes that appear to evolve under positive Darwinian selection. We analyzed expression differences between the ecotypes for selected candidate genes in the GRAS transcription factor family, the chalcone synthase family and in genes involved in cell wall modification that are potentially related to phenotypic differences. We confirm that Reute is a *P. patens* ecotype, and suggest its use for reverse-genetics studies that involve progression through the life cycle and multiple generations.**

**Keywords: *Physcomitrella patens*, ecotype, Reute, sporophyte, microarray, single-nucleotide polymorphism, spore.**

## INTRODUCTION

### The model moss

The moss *Physcomitrella patens* belongs to the Funariaceae with type species *Funaria hygrometrica*, which has been used for physiological studies for more than half a century (Bryan, 1957; Krupa, 1967). Whereas *P. patens* has been used for similar studies starting nearly as long ago (Engel, 1968), the last decade has seen the completion of the nuclear genome sequence (Rensing *et al.*, 2008) and the development of a plethora of tools for this organism (Reski and Cove, 2004; Frank *et al.*, 2005; Quatrano *et al.*, 2007; Kamisugi *et al.*, 2008; Lang *et al.*, 2008; Prigge and Bezanilla, 2010). Today, *P. patens* is one of the primary plant models for evolutionary developmental and

comparative studies (e.g. Mosquna *et al.*, 2009; Okano *et al.*, 2009; Khandelwal *et al.*, 2010; Sakakibara *et al.*, 2013; Horst *et al.*, 2016), and is also employed to study physiology, genome evolution and homologous recombination (e.g. Rensing *et al.*, 2012; Beike *et al.*, 2014, 2015; Charlot *et al.*, 2014).

**Worldwide accessions**

*Physcomitrella* has been described to occur in North America, Europe, Africa, China, Japan and Australia (Frey *et al.*, 2009), and is distributed in the land masses of the Holarctic (Medina *et al.*, 2015). In total, 20 *P. patens* accessions, four *Physcomitrella magdalenae* accessions and 15 *Physcomitrella readeri* accessions have been described, and accessions from all these locations have been cultured axenically *in vitro* (von Stackelberg *et al.*, 2006; Beike *et al.*, 2010, 2014; McDaniel *et al.*, 2010; Medina *et al.*, 2015). The single spore isolated near Gransden Wood (Cambridge, UK) by Whitehouse in 1962 was used initially for *in vitro* culture (Engel, 1968), and became the worldwide laboratory strain *P. patens* Gransden, the genome of which was sequenced by Rensing *et al.* (2008). In addition, the genetically divergent (von Stackelberg *et al.*, 2006) isolate Villersexel K3 (Haute Saône, France; collected by Lüth 2003) was used to generate a genetic map through crossing with Gransden (Kamisugi *et al.*, 2008). The accession Reute was collected by Lüth in 2006 close to Freiburg im Breisgau, Germany, from a moist, disturbed field. Its marker-based genetic distance to Gransden is less than that of Villersexel, and all three accessions can be crossed with each other (von Stackelberg *et al.*, 2006; McDaniel *et al.*, 2010; Perroud *et al.*, 2011; Beike *et al.*, 2014). Such crosses produce viable offspring and have been successfully used to generate a genetic map (Kamisugi *et al.*, 2008) and in forward genetics (Stevenson *et al.*, 2016).

**Sexual reproduction and life cycle**

The induction of *P. patens* gametangia development by low temperature is well established, with incubation at 17°C leading to gametangia development within 7–14 days (Engel, 1968; Nakosteen and Hughes, 1978). The additional shortening of day length and reduction in light intensity further increases the frequency of gametangia (female archegonia, male antheridia) formation in Gransden, with optimal laboratory induction conditions being 15°C, an 8-h photoperiod and 20 μmol $m^{-2}$ $s^{-1}$ (Hohe *et al.*, 2002), mimicking a spring or autumn day. Buds were formed 5–7 days after spore germination, gametophores were formed after 11–13 days, and mature spore capsules were formed 21–28 days after induction on agar (Nakosteen and Hughes, 1978; Ortiz-Ramirez *et al.*, 2015). As antheridia mature earlier than archegonia (Nakosteen and Hughes, 1978), selfing will occur if fertilization by sperm from other plants has failed. *Physcomitrella patens* is known to show a low rate of out-crossing and is predominantly self-fertilising (Perroud *et al.*, 2011), but the haploid moss is able to efficiently purge deleterious mutations (Szovenyi *et al.*, 2014).

Bryophyte spores can survive decades in mud or herbaria (Glime, 2007). For example, spore viability in *F. hygrometrica* was demonstrated after 11 years (Hoffmann, 1970). Moss spore germination can be induced by light, but the quality of the light and the quantity necessary for spore germination is likely to depend on the habitat, e.g. under a canopy versus out in the open, whereas the optimum germination temperature can vary in populations of the same species (Glime, 2007). Spore germination in *P. patens* appears to be suppressed by ultraviolet B (UV-B) irradiation in a dose-dependent manner (Wolf *et al.*, 2010), is completely inhibited by a pulse of far-red light (Possart and Hiltbrunner, 2013) or elevated temperature (Vesty *et al.*, 2016), and depends on phytohormone regulation (Vesty *et al.*, 2016). On agar, spores usually germinate within 3 days (Nakosteen and Hughes, 1978). *Physcomitrella patens* spores are 30 μm in diameter, and around 8000–16 000 are contained per capsule (Nakosteen and Hughes, 1978).

**Transcriptome analyses**

Numerous transcriptomic analyses have been performed, using microarrays based on annotation v1.1 or earlier to analyse ABA, drought stress responses and sporophyte development (Cuming *et al.*, 2007; Komatsu *et al.*, 2009; O'Donoghue *et al.*, 2013), followed by a design based on v1.2 analysing different abiotic stresses and developmental stages (Wolf *et al.*, 2010; Busch *et al.*, 2013; Hiss *et al.*, 2014; Beike *et al.*, 2015). More recently, based on v1.6 (Zimmer *et al.*, 2013), array analyses looked into developmental progression and mutants (Ortiz-Ramirez *et al.*, 2015; Yaari *et al.*, 2015). Although the array designs were based on Gransden, for which the complete genome sequence is available (Rensing *et al.*, 2008), the Villersexel ecotype was successfully hybridized to microarrays (O'Donoghue *et al.*, 2013). With the advances in next-generation sequencing technologies, high-throughput cDNA sequencing (RNA-seq) is now frequently used to measure expression strength and to detect differentially expressed genes (DEGs). In *P. patens*, RNA-seq studies (Table S1) have described the development from protoplasts to protonema, and further on to gametophores (Xiao *et al.*, 2011, 2012), but not yet the development of sporophytes. Recently, the flagship Gene Atlas initiative by the U.S. Department of Energy (DoE), has undertaken deep RNA-seq sequencing to cover the most common developmental stages and perturbations. The moss *P. patens* is one of seven plant 'flagship' organisms (http://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/) tackled in the Gene Atlas project.

In this study we have compared the sexual reproduction of the Reute ecotype with that of Gransden and

Villersexel. We provide and compare transcriptome data of sporophyte development, and have performed comparative molecular variant analyses on the three ecotypes. Thus we introduce the Reute ecotype for reverse-genetics studies that involve progression through the life cycle, overcoming problems that many labs face in using Gransden for such studies.

## RESULTS AND DISCUSSION

### Reute natural conditions

The accession Reute was collected close to Freiburg, Germany, by Lüth in 2006 (Figure S1a). In November 2006 and in October 2008 mature sporophytes were found on a field that is ploughed in autumn, typically in September or October, probably thus exposing spore capsules resting in the soil (Beike *et al.*, 2014). Wet conditions (residual water in furrows) combined with shortened day length induce spore germination and subsequently sporophyte development. Average temperatures 5 cm above the soil were found to be $1 \pm 3$°C from October to December in the years 2006–2015, and in some cases temperatures below 0°C were recorded causing frost coverage (Figure S1b). Daily rainfall was between 0 and 6 mm (Figure S1c), with an average of 2.3 mm per day or 71.9 mm per month. As the field is not shaded, direct sunlight reaches the site (Figure S1b): the light fluence rate in November was measured at >100 μmol $m^{-2}$ $s^{-1}$ (with cloud cover) and >700 μmol $m^{-2}$ $s^{-1}$ (without cloud cover). The natural conditions of growth and sporophyte/spore development of Reute are thus cold temperatures slightly above freezing, wet environment and short days, albeit with direct sunlight, and thus higher light fluence, including UV-B, than a forest floor moss would experience, for example. Noteably, Gransden (although collected at a similar site from furrows of a ploughed field) would experience different weather conditions than Reute, with both less rain and less sun in October at Gransden Wood than at Reute, for example (Figure S2; Table S2). Although the high light conditions found at the Reute site differ from those used to promote sexual reproduction in the laboratory, it should be noted that day length and lower temperature have a larger impact than light fluence rate (Hohe *et al.*, 2002), which is in line with the weather conditions as found at the Reute site in autumn.

### Gametangia and sporophyte development

A number of laboratories working with the sequenced 'Gransden 2004' strain (Rensing *et al.*, 2008) observed a low rate of sporophyte production (Ashton and Raju, 2000; Landberg *et al.*, 2013) and instead used Villersexel to study genome-wide expression patterns during sporophyte development (Landberg *et al.*, 2013; O'Donoghue *et al.*, 2013); however, gametangia and sporophytes do not show any evident morphological differences among Reute, Gransden and Villersexel in axenic *in vitro* culture (gametophore/sporophyte $n = 4908/3700$ for Reute, 2965/191 for Gransden and 1529/1137 for Villersexel).

Archegonia of all three ecotypes (Figure 1) comprise a flask shaped egg cell-containing venter (bracket) and tubular neck canal cells (nc) with no visible aberration. Also, antheridia of these three ecotypes show no visible distinctions, growing in septate bundles at the apex of the gametophore (Figure 1a–c, lower panels). Sporophyte development occurs as described for Gransden (Sakakibara *et al.*, 2008).

The time point at which mature gametangia can be found at the gametophore apex in Reute does not differ from Gransden and Villersexel. Under the growth conditions applied here, most apices of all ecotypes carry mature gametangia at day 19–21 after transfer to short-day conditions. If fully developed gametangia are present, watering leads to synchronized fertilization, as the flagellated antherozoids (spermatozoids/sperm cells) need liquid water in order to swim to the archegonia. At 7–10 days after watering (7–10 daw) a pre-meiotic, elongated and inflated early sporophyte (Figure 1e, ES) is observed. At 9–15 daw spore mother cells (SMCs) are formed in a now spherical, translucent green sporophyte (Figure 1e, PM). SMCs undergo meiosis at around 14–15 daw, and the sporophyte turns from translucent to opaque green (Figure 1e, M). During the ripening process the sporophyte turns from yellow (Figure 1e, Y) to light and dark brown (Figure 1e, LB/B), until mature spores are present around 30 daw. Spores start to develop at an early stage when the cells are still surrounded by a cytoplasmic membrane (Figure 1e, Y), culminating in spore coat-covered mature spores (Figure 1e, B).

Reute demonstrates a sporophyte development frequency that is comparable with Villersexel (Figure 1d; median sporophytes per gametophore in Villersexel, 77%, and Reute, 77%), and is therefore well suited for studies focusing on sexual reproduction, fertilization, embryogenesis and sporophyte development. The sporophyte development frequency of Reute versus Gransden and Villersexel versus Gransden differs significantly, however (Wilcoxon rank tests, $P < 0.01$; median/mean sporophytes per gametophore Reute, 77/80%, Villersexel, 77/80%, and Gransden, 5/6%). While gametangia and sporophytes show no morphological differences and develop at the same rate in all ecotypes, Gransden clearly differs significantly in the number of developed sporophytes compared with Reute and Villersexel. Such phenotypic differences, along with their distinct geographic location and their ability to interbreed are hallmarks of distinct ecotypes.

There are several possible explanations for the observed differences. There may be genetic or epigenetic differences between Gransden and Villersexel that account for this. As

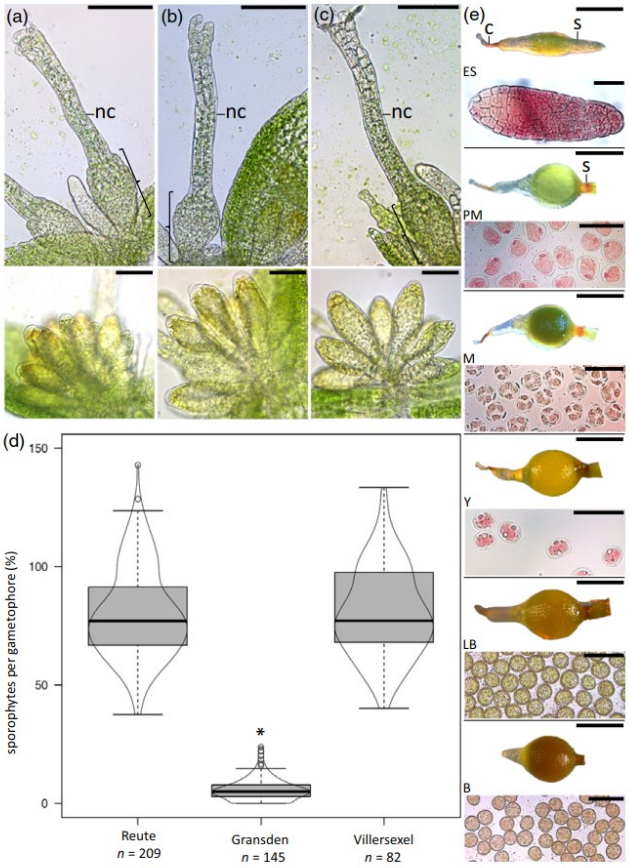



**Figure 1.** Gametangia and sporophyte development. Gametangia (female archegonia and male antheridia) of *Physcomitrella patens* ecotypes Reute (a), Gransden (b) and Villersexel (c). Upper panel shows mature archegonia, consisting of flask-shaped egg-containing venter (bracket) and tubular neck canal cells (nc). Scale bars: 100 μm. Lower panel shows antheridia, occurring in septate bundles at the gametophore apex. Scale bars: 50 μm. (d) Box plot of average number of sporophytes per gametophore (*n* = number of plants) as a percentage. The plot is median-centred, with the grey box representing 50% of the measurements. The whiskers end with the last value in the 1.5 interquartile range (IQR). The edged area shows the distribution of measurements. Sporophyte development of Reute median: 77% versus Gransden 5%. Sporophyte development of Villersexel median: 77% versus Gransden 5%. Differences are significant (Wilcoxon rank test, $P < 0.01$; marked by asterisk), whereas Reute and Villersexel show a comparable proportion of sporophytes (Wilcoxon rank test, $P = 0.84$). (e) Reute sporophyte developmental stages (scale bars: 500 μm), with corresponding spore stages, stained with acetocarmine (scale bars: 50 μm). ES (early sporophyte): elongated premeiotic sporophyte with calyptra (C) and developing seta (S). PM (premeiotic): spherically shaped premeiotic translucent green sporophyte containing spore mother cells, with developed seta (S, brown area). M (meiotic): postmeiotic opaque green sporophyte, cellular content shows spore mother cells with tetrads after metaphase II of meiosis. Y: yellow sporophyte, including ripening spore mother cells. LB: light-brown sporophyte with spores surrounded by a visible spore coat. B: mature brown sporophyte without calyptra, containing mature spores.

only the frequency of sporophyte development appears different there could be a failure of fertilization: either spermatozoids might not be released from Gransden antheridia or they might be less motile. Archegonial development might be affected: during ripening, the archegonial tip cell and inner canal cells degrade (Landberg *et al.*, 2013) to free the way for the entering spermatozoids, a developmental step that might be disrupted. Fertilization or early development of the zygote could be aberrant. A defect during later embryogenesis can be excluded, as no late-stage aborted embryos or sporophytes could be observed. Although the nature of the difference is not the focus of this study, future research might point out why the frequency of sporophyte development differs.

It was suggested that prolonged vegetative cultivation may be the cause for a loss in fertility (Ashton and Raju, 2000), but conditions for sporophyte induction (vessels, substrates, light, temperature) do vary between labs. The conditions used here are adopted from those originally established for Gransden (Hohe *et al.*, 2002).

### Analysis of Reute gametangia and sporophyte development

To address the possibilities described above, we conducted a more detailed developmental analysis. For the Reute ecotype, as is generally known for *P. patens* (Landberg *et al.*, 2013), immature and mature archegonia can be distinguished via the opening of the archegonial tip (Figure 2a,i). This enables spermatozoids to enter and reach the egg cell in the archegonial venter (Figure 2d, arrowhead). During growth and ripening, antheridia undergo an increase of cell size and change color from green (immature; Figure 2b,e) via yellow (mature) to brown (post release; Figure 2c,h). Water is required for fertilization to

occur, not only as the transport medium for the flagellated spermatozoids, but also for their release, as the water is taken up by the antheridial tip cells, causing them to swell and finally burst to release spermatozoids (Figure 2c,h, arrowheads). Figure 2(e) shows an early antheridial stage, at which anticlinal cell division can be observed. In the mature archegonium the elongated outer neck canal cells can be seen clearly (Figure 2f, light green, arrowhead), as well as a paraphysis (a sterile organ consisting of elongated cells with a swollen apical cell; Figure 2f,i, blue). In Figure 2(g) a sporophyte with detached calyptra is shown from the top, and the distribution of the outer sporophytic cells can be observed with several cells having recently divided. In opened sporophytes the cellular content can be observed (Figure 2j, ochre). In summary, the detailed analysis of Reute gametangia and sporophyte development confirms its similarity with that of Gransden, including the presence of paraphyses and details of archegonial/antheridial growth (Landberg *et al.*, 2013).

**Expression differences between Gransden and Reute gametophores**

To determine whether – despite similar development – there are differences in gene expression, we analysed expression profiles. To find differences in gene expression between Gransden and Reute, Hiss *et al.* (2014) performed a whole transcriptome microarray analysis of gametophores with developed gametangia (adult gametophores) for both ecotypes. Here, we analyze these data and find 262 DEGs (Appendix S2), 250 of which were found to show lower and 12 of which were found to show higher expression in the Reute ecotype, when compared with Gransden. Of particular interest are transcription factors (TFs) and transcriptional regulators (TRs) that are differentially expressed between the two ecotypes, as these may underlie phenotypic differences such as the sporophyte development frequency. We find 10 such proteins (Table 1), with PPM3 (Pp3c14_22180V3.1), a member of the moss-specific MADS-box containing subfamily MIKC* (Barker and Ashton, 2013), among them. Members of this subfamily regulate pollen development in *Arabidopsis thaliana* (Gramzow and Theissen, 2010), and therefore could be involved in the development of moss spores, which represent a developmentally analogous structure (Brown and Lemmon, 2011; Daku *et al.*, 2016; Vesty *et al.*, 2016).

We also find a member of the GRAS TF family among the DEGs. The proteins of this family share the GRAS DNA-binding domain (Li *et al.*, 2016), whereas the N-

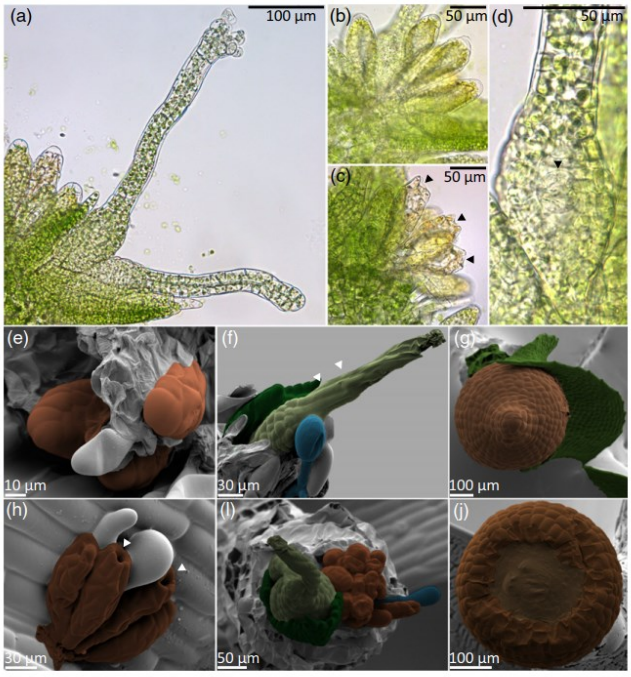



**Figure 2.** Gametangia and sporophytes of *Physcomitrella patens* ecotype Reute. (a) Immature (closed tip) and mature (open tip) archegonia. Immature antheridia (b) and mature antheridia (c), which release spermatozoids through the burst tip cells (arrowheads). (d) Unfertilized egg cell in archegonial venter (arrowhead). (e, f) False-colored cryo-SEM images: (e) early antheridia showing anticlinal cell division; (f) mature archegonium (light green) with elongated outer neck cells (arrowhead), paraphyse (blue) and young phyllid (green). (g) Mature sporophyte, top view (sporophyte brown, phyllid green). (h) Mature antheridia after spermatozoid release through burst tip cell (arrowheads). (i) Gametophore apex, top view with young phyllid (green), mature archegonia (light green), antheridia (brown) and paraphysis (blue). (j) Sporophyte opened at the tip, showing cellular content (ochre).

**Table 1** ID numbers and annotation of 10 transcription factors or transcriptional regulators expressed at a lower level in Reute, as compared with Gransden

| CGI v3 | TF/TR family | Fold change |
|---|---|---|
| Pp3c14_22180V3.1 | MADS | 4.6 |
| Pp3c13_3830V3.1 | AP2/EREBP | 6.7 |
| Pp3c15_3180V3.1 | C2C2_Dof | 4.7 |
| Pp3c1_35770V3.1 | Argonaute | 5.2 |
| Pp3c8_230V3.1 | Zinc finger, AN1 and A20 type | 5.0 |
| Pp3c4_12970V3.1 | ARF | 6.3 |
| Pp3c11_21140V3.1 | GARP_G2-like | 4.4 |
| Pp3c2_20930V3.1 | GRAS | 4.2 |
| Pp3c2_8880V3.1 | bHLH | 4.6 |
| Pp3c11_20170V3.1 | bHLH | 8.5 |

Based on Combimatrix microarray data comparison between adult gametophores (bearing gametangia) of the Gransden and Reute ecotypes.

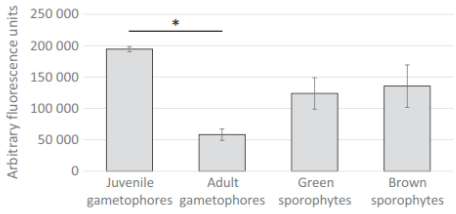



**Figure 3.** Bar chart of GRAS Pp3c2_20930V3.1 microarray expression values. Four different *Physcomitrella patens* Reute developmental stages for the gene *Pp3c2_20930V3.1* are shown based on novel v1.6 NimbleGen microarray data. Error bars show standard deviations for two or three biological replicates. *Significant difference.

terminal parts are predicted to contain molecular recognition features (Morfs) that are important for protein–protein interactions (Sun *et al.*, 2011). GRAS family members are responsible for regulating different plant growth and development steps (Bolle, 2004), and some activate meiosis-specific genes (Morohashi *et al.*, 2003). The intron-less GRAS gene Pp3c2_20930V3.1, encoding a 770 amino acid protein, is expressed at a fourfold lower level in Reute adult gametophores than in Gransden (Figure S3; Table 1). In the Gransden ecotype, Pp3c2_20930V3.1 is more strongly expressed in protoplasts and under UV-B treatment (Figure S4), but does not show reduced expression during the development of sexual organs (i.e. in adult gametophores). Variation in expression can also be seen in the novel Reute array data presented here (Figure 3). Interestingly, the pronounced decrease in expression of this gene in Reute during the development of sexual organs is not visible to this extent in Gransden. Differences between the ecotypes with regards to the frequency of sporophyte development might thus be associated with this GRAS TFs.

In summary, 10 TFs/TRs are differentially regulated between the ecotypes and are good candidates for investigating the differences in frequency of sporophyte development.

### Expression profiling of Reute developmental stages

Upstream regulators such as TFs/TRs often control downstream effector genes that execute the actual phenotypic alterations: here, we focus on such output genes. As we were interested in the sporophyte development of Reute, we generated NimbleGen microarray data for different stages of development: gametophores without gametangia (juvenile), with gametangia (adult), and green sporophytes (developing, pre-meiotic; PM in Figure 2) as well as brown sporophytes (mature, post-meiotic; B in Figure 2). DEGs were computed in pairwise fashion along the developmental progression (Figure 4). We find a high number of DEGs between juvenile and adult gametophores (6021), and also between adult gametophores and green sporophytes (2492). Between green and brown sporophytes we find 313 DEGs.

We focused on genes showing differential expression between adult gametophores and green sporophytes in the Reute ecotype, as this developmental step seems to be affected in the Gransden ecotype. We specifically examined only genes showing differences between the Gransden and Reute ecotypes. We identified 41 genes, 36 of which are expressed at a lower level in the Reute ecotype relative to Gransden (Appendix S2). For selected genes we confirmed the expression profile during sporophyte development by qPCR (Figure 5).

Two of the 10 differentially expressed TFs/TRs between Gransden and Reute (MADS, Pp3c14_22180V3.1; AP2/EREBP, Pp3c13_3830V3.1) also show differential expression during sporophyte development. Aside from that, the list contains genes that are predicted to code for cell wall-modifying enzymes like pectin methylesterase (Pp3c5_23400V3.1) and xylosyltransferase (Pp3c23_380V3.1). Both genes show a stronger expression in adult gametophores than in green and brown sporophytes, suggesting that their products are more active during the late gametophytic c stage. As the development of the sporophyte involves cell wall restructuring (O'Donoghue *et al.*, 2013), the observed expression differences may contribute to the phenotypic differences.

We further find three DEGs that belong to the chalcone synthase (CHS) gene family (Figure S4), namely Pp3c2_32400V3.1 (CHS1a), Pp3c2_32960V3.2 (CHS10 PpASCL; Figure 6) and Pp3c11_2990V1.1 (CHS4; see Table S5 for ID). Chalcone synthases (CHS) catalyze one of the first steps of flavonoid biosynthesis, and are encoded by an expanded gene family in *P. patens* (Koduri *et al.*, 2010; Wolf *et al.*, 2010). The CHS genes Pp3c2_30620V3.1 (CHS01), Pp3c2_32400V3.1 (CHS1a) and Pp3c2_32320V3.1

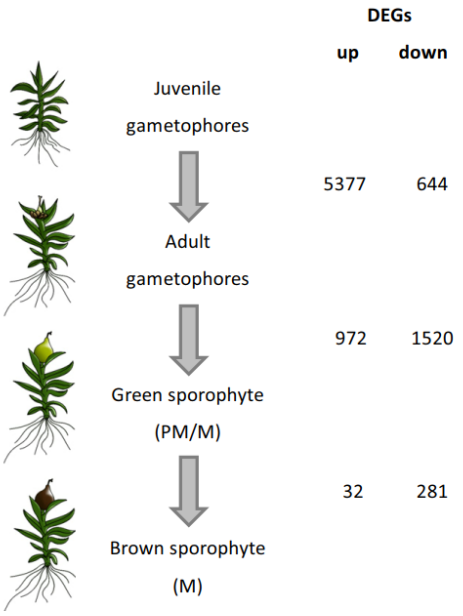



**Figure 4.** Scheme of *Physcomitrella patens* developmental stages and differential gene expression. Left, developmental stages for which NimbleGen microarray data were generated (Reute ecotype). Gametophores were harvested without rhizoids, PM/M and M sporophytes (cf. Figure 2) were separated from the gametophores at harvest. Right, differentially expressed gene (DEG) summary, showing the number of significant DEGs (CyberT test, Benjamini–Hochberg corrected $P < 0.05$; see Experimental procedures for details) that are expressed at higher or lower level between the developmental stages shown, based on the NimbleGen microarray data.

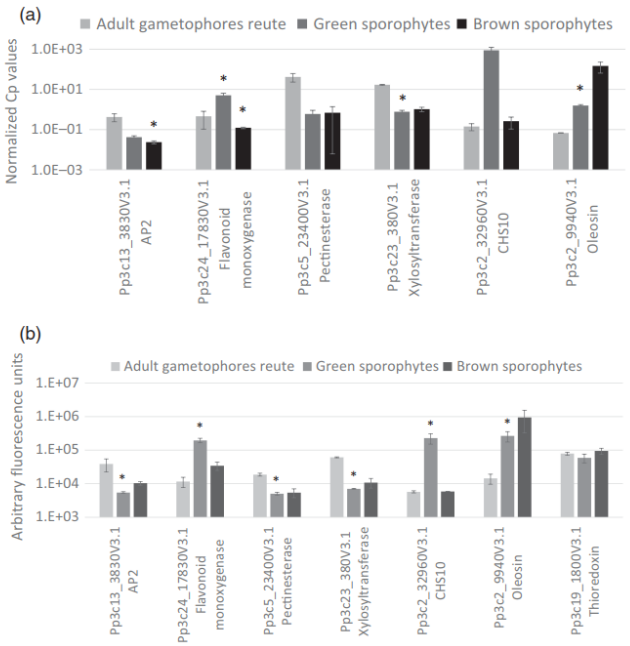
(CHS1c; cf. Figure S6) were found to be induced after UV-B treatment, and are suggested to function as a molecular sunscreen (Wolf *et al.*, 2010); all three genes are less expressed under drought (Stevenson *et al.*, 2016). Here, we find CHS1a to be repressed in brown sporophytes, with a higher expression in all other stages analyzed, namely juvenile gametophores, adult gametophores and green sporophytes (Figure 6). In contrast to CHS1a, CHS10 is induced in green sporophytes as compared with the other three developmental stages, which is also supported by the qPCR analysis (Figure 5). This gene is a functional ortholog of the *A. thaliana* type-III polyketide synthase A (PKSA) belonging to the anther-specific chalcone synthase-like (ASCL) genes, and has been shown to be part of the sporopollenin biosynthesis pathway in *P. patens* (Colpitts *et al.*, 2011; Daku *et al.*, 2016). The repression of the UV-B induced CHS1a/Pp3c2_32400V3.1 accompanied by the induction of the spore coat formation gene Pp3c2_32960V3.2 suggests that biosynthesis of UV-B absorbing quercetin and related flavonoids (Wolf *et al.*, 2010) is no longer needed once sporopollenin is formed. At the same time, the lower expression of CHS1a might be associated with dehydration of the maturing sporophyte.

### Genetic variation among Villersexel, Reute and Gransden

To determine the potential genetic basis for the observed expression and phenotypic differences, we analysed genetic variation of the ecotypes using novel Reute genomic DNA data. Reute and Villersexel can be crossed with each other and with Gransden; however, based on selected markers the genetic distance between Villersexel and Gransden is much greater than between Reute and Gransden (McDaniel *et al.*, 2010). Among different European accessions Villersexel appears most genetically divergent from Gransden (Kamisugi *et al.*, 2008). With the Reute ecotype we present an alternative ecotype with a genetically closer Gransden, yet suitable for both 'forward' (map-based) and 'reverse' genetics approaches. The lower number of polymorphisms makes reverse-genetics approaches based on the Gransden reference genome easier. We sequenced genomic DNA from Reute gametophores to assess the precise genetic distance by evaluating all single-nucleotide polymorphisms (SNPs) and indels. Comparison of Reute genomic DNA (gDNA) sequence data with the Gransden reference genome identified 264 782 SNPs and 16 292 indels (7874 insertions; 8418 deletions), resulting in a polymorphism density of one SNP every 1783 bases and one indel every 28 857 bases. For Villersexel we find 2 497 294 SNPs and 172 833 indels (77 522 insertions; 95 311 deletions), resulting in a density of one SNP every 188 bases and one indel every 2724 bases. SNP densities between *A. thaliana* ecotypes have been shown to occur between one SNP per 149 bp and one SNP per 285 bp (Cao *et al.*, 2011), similar to the densities found in Villersexel, which is surprising given that the rate of mutation fixation is lower in *P. patens* (Rensing *et al.*, 2007). Reute exhibits an almost 10-fold lower SNP density than Villersexel, although the two collection sites are only 100 km apart geographically, with a negligible difference in latitude, but separated by a mountain range. Reute is found on a field that is regularly plowed in autumn, and Villersexel is found at a dried fish pond, a location that is also regularly flooded. Environmental/microclimatic differences at the two sites might differ, however. The discrepancy between genetic and geographical distance may be explained by the distribution of *P. patens* spores via migrating birds that has been proposed recently (Beike *et al.*, 2014).

We find that most SNPs and indels fall into intergenic regions (about 80%, see Table 2), and into the 2000 bp upstream and downstream of the transcript, consisting of untranslated regions (UTRs) and potential promoter areas (about 15%). Out of the 35 302 genes predicted, about half

**Figure 5.** Bar chart of expression values derived from qPCR (a) and microarray (b) analysis. qPCR expression values are normalized to the reference gene *Pp3c19_1800V3.1*, which shows a steady expression in the microarray data across the measured tissues. *Significant changes, compared with the preceding developmental stage.



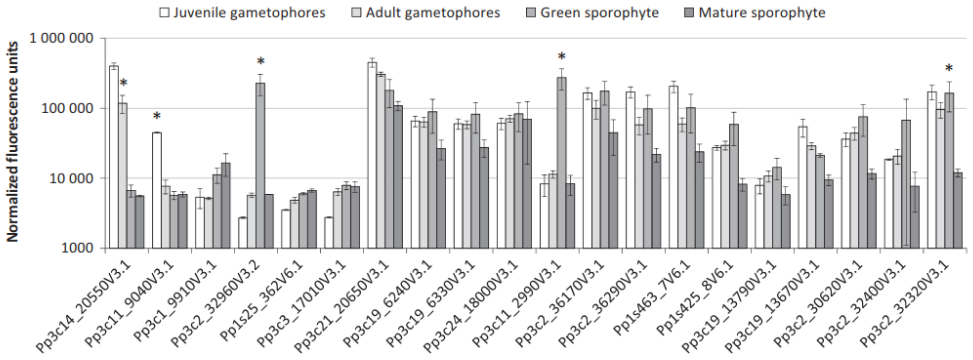




**Figure 6.** Bar chart of microarray expression values (arbitrary units) for 20 chalcone synthase (CHS) genes. Four *Physcomitrella patens* Reute developmental stages are shown, with novel data based on the v1.6 NimbleGen array (corresponding v3.3 gene IDs are shown, except for cases where no v3.3 model was available). Error bars indicate standard deviations of two or three biological replicates. *Significant changes, compared with the following developmental stage (FDR; corrected CyberT test $P < 0.05$).

(15 178) had an SNP within their gene body or promoter in Reute, and 5842 had an indel. In Villersexel almost all (32 473) genes contained an SNP and 27 722 contained an indel. Most SNPs in Reute as well as in Villersexel cause a mis-sense (63%) or a silent (35%) mutation in the coding sequence, and only a few cause a non-sense mutation (2%). Premature stop codons were introduced into 74 genes in Reute (Table S5a) versus 602 genes in Villersexel

by SNPs, and into six genes in Reute (Table S5b) versus 42 genes in Villersexel by an indel. We find 24 genes in Reute and 190 genes in Villersexel that are longer than in Gransden, and therefore may contain a premature stop codon in Gransden.

The Reute SNPs and indels are unevenly distributed on the chromosomes, with several areas of higher SNP density being evident (Figure 7). We selected areas that show significantly higher SNP density (false discovery rate (FDR) corrected $P < 0.01$, see Experimental procedures for details; Table S6), and chose the longest two regions, located on chromosomes 8 and 19, for closer inspection. Within the 1-Mbp peak region on chromosome 8 we find 21 gene models, three of which contain SNPs in their coding sequence; for the 1.7-Mbp peak on chromosome 19 we find 72 gene models, 28 with an SNP in their coding sequence (CDS). For six gene models of the chromosome-8 peak and 11 gene models of the chromosome-19 peak we find close paralogs (BLAST hits with ≥90% identity and length ≥90 aa), often on the same chromosome. Such paralogs, some of them tandemly arrayed genes, can help to provide higher gene-product dosage and might be subject to concerted evolution by gene conversion, in which one copy is 'overwritten' by homologous recombination using the other copy as a template (Wang and Paterson, 2011; Wang *et al.*, 2011). Interestingly, in Reute as well as in Gransden we find three CHS pairs that show identical protein sequences and are located close to their respective partner on chromosomes 2 and 19 (Pp3c2_32400V3.3/CHS1a and Pp3c2_32320V3.3/CHS1c, Pp3c2_36170V3.3/CHS2b.1 and Pp3c2_36290V3.3/CHS2c, Pp3c19_6320V3.3/CHS3.1 and Pp3c19_6330V3.3/CHS3.3), potentially providing higher gene dosage via concerted evolution.

Gene ontology (GO) bias analysis of the genes in the peak regions versus all genes finds diverse GO terms over-represented in the 21 gene models from chromosome 8, e.g. thioredoxin biosynthesis, mRNA processing and superoxide responses. The 72 gene models on chromosome 19 show an over-representation of genes, e.g. involved in cyanate biosynthesis and mitochondrial electron transport (for a full list of GO terms and gene IDs see Tables S7 and S8). Hence, many of the genes in the two SNP hot spots are potentially involved in radical scavenging. This could be an adaptation to environmental conditions that are characterized by higher levels of photonic radiation, as a result of the more southerly latitude and less average cloud cover.

The late embryogenesis abundant 1 (*LEA1*) gene Pp3c22_8970V3.2 contains a premature stop codon in the first exon. It is expressed, but the gene product does not seem to be necessary for normal growth and development (Kamisugi and Cuming, 2005). The SNP causing the premature stop codon is present in Reute, but not in Villersexel, where a CAG is present, demonstrating again that Reute is genetically closer to Gransden than Villersexel, and that genetic variation of this particular gene varies among ecotypes.

### Reute pseudoreference genome and genes under selection

In order to determine genes under selection, and to make Reute more useful for the community, Reute SNP and deletion data were incorporated into the Gransden reference genome sequence to create a 'pseudogenomic sequence' that can be used for Basic Local Alignment Search Tool (BLAST) searches, phylogenetic analysis or planning of reverse-genetics experiments (available at http://plantco.de/ReutePseudogenome.fa). From the pseudogenome we calculated the ratio of non-synonymous/synonymous ($K_a/K_s$) substitutions between Gransden and Reute, which was possible for 320 genes, and focused on the top and bottom 5% (16 genes each). In the top 5% we found 15 genes showing a $K_a/K_s$ ratio larger than 2 (Table S9), and

**Table 2** Summary table of SNP effect analysis for *Physcomitrella patens* ecotypes Reute and Villersexel, as compared with Gransden

| | Reute – SNPs | | Reute – Indels | | Villersexel – SNPs | | Villersexel – Indels | |
|---|---|---|---|---|---|---|---|---|
| Type | Count | Percent | Count | Percent | Count | Percent | Count | Percent |
| Intergenic | 249 609 | 82.6 | 13 560 | 59.0 | 2 355 783 | 81.1 | 155 247 | 63.0 |
| Upstream | 18 628 | 6.17 | 3729 | 16.2 | 200 512 | 6.91 | 39 690 | 16.1 |
| Downstream | 17 753 | 5.88 | 2835 | 12.3 | 196 593 | 6.77 | 32 902 | 13.4 |
| Intron | 5132 | 1.70 | 1139 | 4.96 | 50 725 | 1.75 | 7868 | 3.19 |
| Exon | 4542 | 1.50 | 381 | 1.66 | 38 405 | 1.32 | 1673 | 0.68 |
| 5′-UTR | 2508 | 0.83 | 507 | 2.21 | 22 546 | 0.78 | 3292 | 1.34 |
| 3′-UTR | 2272 | 0.75 | 487 | 2.12 | 22 994 | 0.79 | 3280 | 1.33 |
| Other | 1683 | 0.56 | 337 | 1.47 | 16 022 | 0.55 | 2356 | 0.96 |
| Sum | 302 127 | | 22 975 | | 2 903 580 | | 246 308 | |

The type column lists the possible locations of SNPs, both upstream and downstream, constituting the 2000-bp regions in front and behind of the transcript sequence and intergenic the region between the genes (excluding up- and downstream regions). Transcript regions were assigned according to the annotation version 3.1 from http://www.cosmoss.org. For each ecotype the count and percentage among all the locations is listed. The last row shows the sum of all effects. The analysis was performed with SnpEff and the type 'Other' summarizes 'none', 'splice site acceptor', 'splice site donor', 'splice site region' and 'transcript'.

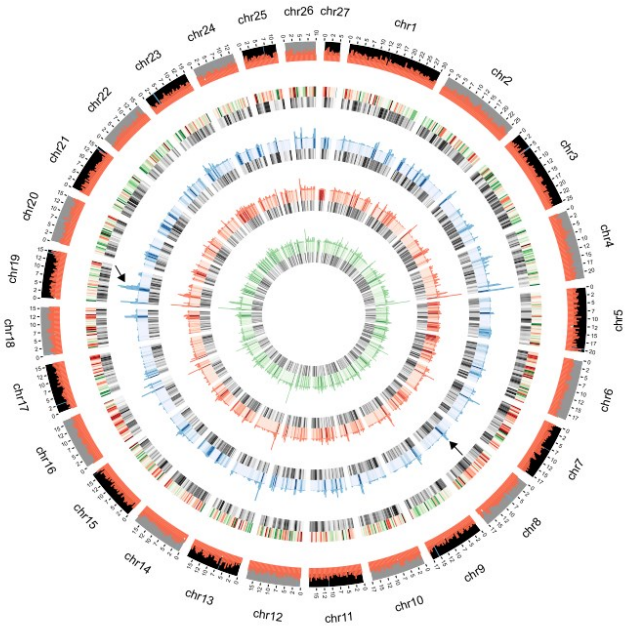



**Figure 7.** Circos plot showing the 27 *P. patens* chromosomes. From outer to inner ring: (i) average gene expression (log2) based on the novel NimbleGen microarray data of four Reute developmental stages (red histogram); (ii) 0–1, normalized gene expression (log2) based on the NimbleGen data, average of four developmental stages (bars in red showing higher expression with values closer to 1 and bars in green showing lower expression with values closer to 0); in the same ring 0–1 normalized gene density is shown (also shown for comparison in the four inner rings, with gray bars of darker color depicting higher density); (iii) SNP density histogram (blue); (iv) deletion density (orange bars); (v) insertion density (green bars). All values were summarized by a sliding window approach with a window size of 500 kbp. Black arrows indicate the two longest high-density SNP peaks on chromosomes 8 and 19.

therefore these genes could be candidates to evolve under positive Darwinian selection.

In conclusion, Reute shows several SNP hot spots that also contain genes with changed coding sequences. Based on the two ecotypes, a number of Reute genes might be subject to positive (Darwinian) selection, which in turn could reflect adaptation to a slightly different niche.

### Determination of a robust set of genes differentially expressed in the *P. patens* sporophyte

Comparison of cross-platform and cross-ecotype data is notoriously difficult. In order to learn whether there is a robust set of genes that are differentially expressed in sporophytes, we analysed all available data sets. Gene expression for green and brown sporophytes from Reute was measured via microarray in this study and via RNA-seq by the JGI Gene Atlas project (http://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/). We compared the DEGs derived from both technologies between green and mature sporophytes (Figure 1e, PM/M and B; Appendix S2). For genes with higher expression in mature sporophytes we found 32 genes via the array and 526 genes via RNA-seq data, 15 of which overlap. For genes with lower expression, the array detects 281 and RNA-seq detects 1030 genes, with an overlap of 129. GO annotation for the 15 more highly expressed genes in the overlap shows over-represented categories including 'reproductive structure development' and 'post-embryonic development' (Figure S5), in accordance with expectations. The fraction of DEGs that overlap between the two platforms is comparable with other studies; in general some technology-specific differences identified by microarray and RNA-seq analyses are common (Marioni *et al.*, 2008; Zhao *et al.*, 2014). Nevertheless, combining results obtained with different technologies results in a high-confidence set of genes involved in the examined developmental stage or perturbation. In summary, we have defined a robust set of DEGs during Reute sporophyte maturation.

### Conclusions

In comparing the *P. patens* ecotypes Gransden, Reute and Villersexel we observe a ~15-fold difference in the number of sporophytes that develop under standardized conditions, with Gransden showing just a few, compared with Reute and Villersexel. The reduced Gransden rate may, however, not represent the natural trait, as Gransden has been cultivated under lab conditions for longer time periods than Reute and Villersexel. We do not observe any gross morphological differences during gametangia and sporophyte development; however, several hundred genes are differentially expressed between Gransden and Reute,

including transcription factors that are known to control developmental processes, members of the CHS family potentially involved in spore (coat) development, and genes encoding cell wall modification enzymes.

We provide a detailed description of gametangia and sporophyte development, and show Reute to be in accordance with previous descriptions of *P. patens*; however, the Reute genome contains hot spots of genetic variation as well as genes under positive selection.

The Reute ecotype thus combines the advantages of high fertility for forward genetics with the capacity for gene targeting of *P. patens*, enabling studies of the whole life cycle. The Reute ecotype is already used by several labs, and has successfully been used for transient and stable transfections. To facilitate these applications we provide a pseudogenomic sequence as well as a set of expression profiling data. Using cross-platform data we define a confident set of 15 genes expressed in mature sporophytes that can be used, for example as expression markers.

## EXPERIMENTAL PROCEDURES

### Plant material

*Physcomitrella patens* ecotypes Gransden (Rensing *et al.*, 2008), Reute and Villersexel (von Stackelberg *et al.*, 2006) were cultivated on solidified [1% (w/v) agar] mineral medium (Knop's medium; Knop, 1868), on 9-cm Petri dishes enclosed by laboratory film, and maintained at 22°C with a 16-h light/8-h dark regime under 70 μmol $m^{-2}$ $s^{-1}$ white light (long-day conditions), as previously described (Hiss *et al.*, 2014). All ecotypes are available at the international moss stock center (IMSC, http://www.moss-stock-center.org) or from the authors upon request. For sporophyte induction, Petri dishes were transferred to 16°C with an 8-h dark/16-h light regime under 20 μmol $m^{-2}$ $s^{-1}$ white light (short-day conditions), as described by Hohe *et al.* (2002), but using medium without supplements. For sporophyte production, 10 gametophores per Petri dish were evenly distributed into the agar and grown under long-day conditions. After moving the plates to short-day conditions, plants were assessed for gametangia appearance and subsequently watered with sterile tap water (Hohe *et al.*, 2002). As fertilization requires water, this procedure ensures a high rate of synchronization of sporophyte development. Fertilization does not regularly occur under the conditions applied here (including eight-fold air exchange per hour and the use of laboratory film to wrap the dishes, allowing gas exchange), as not much condensing water drops onto the gametophores.

### Counting sporophytes per gametophore and statistical analysis

To determine sporophyte development rates, a minimum of five replicate plates were set up as described above for each ecotype. After a minimum of 30 days after watering (30 daw), sporophytes per gametophore were counted (all developmental stages that could be clearly determined as a sporophyte were taken into account) and summarized per plant. Wilcoxon rank tests were performed in R (R Development Core Team, 2008) using the function 'Wilcox.test'. Box plots with added distribution of measurements were generated in R using the function 'boxplot' and the additional package 'caroline', with its function 'violins' (Schruth, 2012).

### Acetocarmine staining

Staining was performed using the method described by Belling (1921). Briefly, acetocarmine staining solution was prepared with 350 ml acetic acid, 650 ml water and 20 g acetocarmine, boiled until completely dissolved, filtered and stored in the dark at room temperature 20-24°C. Sporophytes were fixed for a minimum of 24 h in ethanol/acetic acid (3:1). To stain, fixed tissue was transferred to a microscope slide and squashed to release sporophyte contents; embryonic content was prepared manually using a binocular microscope and forceps. A few drops of staining solution were added and stained for 10 min before image analysis.

### Microscopic imaging of gametangia and sporophytes

The preparation of gametangia was performed using a binocular microscope (S8Apo with MC170HD camera; Leica, http://www.leica.com). Microscopic images were taken with an upright DM6000 microscope (Leica; camera DFC295). Microscopy pictures were processed using Photoshop CC (Adobe Systems Software Ireland Ltd). The brightness and contrast of light microscopy pictures was adjusted, and cryo-SEM images were false colored for enhanced visibility.

### Cryo-SEM analysis of gametangia and sporophyte

For analysis a Philips XL30 ESEM with Cryo Preparation Unit Gatan Alto 2500 was used. Prepared plant material was applied to the specimen holder with freeze-hardening glue and biological samples were preserved by fast-freezing in liquid nitrogen. Afterwards the specimen holder was inserted into the sputter chamber and coated with gold.

### DNA isolation

Genomic DNA was isolated from plant material according to a modified Dellaporta protocol (Dellaporta *et al.*, 1983). After the isopropanol precipitation the dry pellet was dissolved in 700 μl TE buffer (pH 8), 1–3 μl RNaseA (10 mg $ml^{-1}$) was added and incubated for 10 min at 37°C. To purify the DNA, 600 μl phenol/chloroform 1:1 was added, mixed, centrifuged at 10 000 *g* for 1 min and the aqueous phase extracted. To this phase 600 μl chloroform/isoamylalcohol 24:1 was added, mixed, centrifuged at 10 000 *g* for 1 min and the aqueous phase extracted. To precipitate the DNA, 70 μl 3 M Na-acetate and 500 μl isopropanol were added, mixed and centrifuged at 10 000 *g* for 10 min. The pellet was washed with 1 ml 70% ethanol, dried and subsequently dissolved in deionized water. Concentration and quality was tested with the Nanodrop 1000 (ThermoFisher Scientific, http://www.thermofisher.com) and by agarose gel electrophoresis.

### RNA isolation

RNA was isolated from plant material using the RNeasy Plant Micro Kit (Qiagen, http://www.qiagen.com), following the manufacturer's instructions. RNA concentration and size distribution was tested on the 2100 Bioanalyzer (Agilent Technologies, http://www.agilent.com) with the Agilent RNA 6000 Nano Kit to determine quantity and quality.

### NGS analysis

Sequencing data from genomic DNA were retrieved from NCBI SRA in the case of *P. patens* accession Villersexel (SRX030894). For the *P. patens* accession Reute genomic DNA was extracted from gametophores and sequenced on one lane of the Illumina

HiSeq 2500 (100-nt paired end) at the Max Planck-Genome-centre Cologne (http://mpgc.mpipz.mpg.de). The library was prepared according to the Illumina TruSeq protocol with an insert size of 300–400 bp. For Villersexel we started with 201 288 783 paired end reads (SRA: SRX030894), and for Reute we started with 150 711 864 paired end reads (SRA: SRX1528135). Read quality and trimming efficiency was evaluated with FASTQC 0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). All sequence data were trimmed with TRIMMOMATIC 0.32 (Bolger *et al.*, 2014) using the following parameters: -phred33 ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:8:5 SLIDINGWINDOW:4:15 TRAILING:15 MINLEN:35.

After trimming we used 144 872 394 paired end reads for Villersexel and 145 604 667 paired end reads for Reute, for mapping.

All mapping steps were performed with GSNAP 2014-10-22 (Wu and Nacu, 2010), with default parameters. Trimmed reads were mapped against the chloroplast genome (NC_005087.1). Mapped reads were removed from the read pool and the same procedure was repeated for the mitochondrial genome (NC_007945.1) and the ribosomal rRNA genes (HM751653.1, X80986.1, X98013.1). After these steps the reads were mapped against the *P. patens* genome assembly V3 (DoE-JGI, http://phytozome.jgi.doe.gov). Duplicate reads were removed with SAMTOOLS RMDUP (Li *et al.*, 2009) to account for potential PCR artifacts. SNP calling was performed with GATK 3.3.0 (McKenna *et al.*, 2010) and SAMTOOLS 0.1.19. Reference and bam files were indexed with SAMTOOLS FAIDX/INDEX. Adding read-group information and sorting the bam file was achieved with PICCARD-TOOLS 1.115 (http://broadinstitute.github.io/picard). GATK was used as recommended by the Broad Institute for species without reference SNP databases, but the last quality recalibration step was omitted, as no large set of confirmed SNPs is available for *P. patens*. The second SNP calling pipeline used MPILEUP, BCFTOOLS and VARFILTER, with default parameters.

Intersections between.vcf files were extracted with BCFTOOLS 1.2 (http://samtools.github.io/bcftools/bcftools.html). Indel events and SNP events were separated with GATK SELECTVARIANTS. Only homozygous SNPs were used for further analysis because *P. patens* is a haploid organism, and only homozygous SNPs are expected.

Variants were annotated with SNPEFF 4.1 g (Cingolani *et al.*, 2012) using the COSMOSS 3.1 (https://www.cosmoss.org/physcome_project/wiki/Genome_Annotation/V3.1) gene annotation.

Depending on the SNP calling tool employed, we found 3–4 million SNPs between Gransden and Villersexel in the unfiltered data. The overlap of the GATK and SAMTOOLS SNP callers contains almost 1.9 million SNPs and 8900 insertions/deletions (indels). To test the sensitivity of our approach we used a set of 4650 SNPs from the Villersexel accession that were confirmed by an SNP bead array (Appendix S1). Of the SNP array probes, 4628 can be mapped to the V3 genome assembly, and each of the SNP callers (GATK, SAMTOOLS) calls >90% of these test sets on the Villersexel data, showing that the approach used is highly sensitive. Throughout the manuscript we use the data set from the GATK SNP calling because GATK has a quality recalibration step and performs realignment around insertions and deletions.

For the ecotype Reute at each SNP position the corresponding reference allele was replaced by the alternative allele, producing pseudo-chromosome sequences. Using the COSMOSS 3.1 gene annotations, coding sequences were extracted for each gene model and codon alignments were generated for Gransden and Reute. Subsequently, synonymous and non-synonymous nucleotide diversity was calculated using the Yang and Nielson method, as implemented in KAKS_CALCULATOR (Wang *et al.*, 2010). For chromosome-wide plots a sliding window approach (window size, 500 kbp; jump size, 400 kbp) was used and nucleotide diversity values were calculated with VARISCAN (Hutter *et al.*, 2006).

### SNP peak detection

Window-wise (50 kbp with 10-kbp overlap), SNP numbers were extracted from the 'pseudogenome' FASTA file by a custom R script. The R functions fisher.test and p.adjust (method = 'hochberg') were used to select fragments that show a significantly (adjusted $P$ value < 0.01) higher SNP number than the chromosome average. An SNP hot spot was called if at least five adjacent fragments showed a significantly higher SNP number.

### Microarray analyses

The NimbleGen 12 × 135k DNA microarray covers 32 851 transcripts, with a total of 130 221 probes in 32 741 probe sets (for 99.66% of the transcripts), of which 353 (1.07%) map redundantly and 87 (0.26%) contain fewer than four probes. Besides the v1.6 transcripts, spikes and negative controls were included, as were 580 v1.2 gene models that lack a v1.6 equivalent but were shown to be differentially expressed based on existing Combimatrix microarray data.

About 200 ng of total RNA was reverse transcribed and amplified using the WTA Kit (Sigma-Aldrich, https://www.sigmaaldrich.com). One microgram of cDNA was labeled with Cy3 according to the NimbleGen One-Color DNA Labeling Kit (Roche, http://www.roche.com); 4 µg of labeled cDNA was used for hybridization on the NimbleGen 12 × 135k DNA microarray, probe design OID33087 (Roche), according to the manufacturers' protocol using the NimbleGen Hybridization Kit (Roche). The NimbleGen Wash Buffer Kit (Roche) was used to prepare the slide for scanning.

The arrays were imaged using a laser scanner Agilent G2565CA Microarray Scanner System (Agilent Technologies). The image of the arrays was cut into single array images using NIMBLESCAN 2.5 (Roche), and the pixel intensities were extracted with the same software.

Microarray expression data were analyzed with ANALYST 7.5 (Genedata, https://www.genedata.com). Median condensed probe set expression values were quantile-normalized and analyzed further, as previously described (Wolf *et al.*, 2010). Box plots, hierarchical clustering (Figures S6, S7) and CyberT test analyses were performed with R (R Development Core Team, 2008).

### Quantitative real-time PCR (qPCR)

For validation of genes found to be differentially expressed in the microarray data, the treatment of *P. patens*, harvesting and RNA extraction was carried out as described above. cDNA was synthesized using the Superscript III kit (Life Technologies, now ThermoFisher Scientific, http://www.thermofisher.com) following the manufacturers' protocol. Real-time qPCR was carried out using the SensiMix SYBR green No-ROX kit (Bioline, http://www.bioline.com) with a 10-µL reaction volume. Primers were designed with PRIMER 3, aiming at an annealing temperature of around 60°C for all primers and 3′ clamp and intron-spanning, where applicable (Koressaar and Remm, 2007; Untergasser *et al.*, 2012), and checked for a single genomic locus via BLAST (cosmoss.org; see Table S3 for primer sequences). Melting curve analysis and non-template controls (NTCs) were carried out routinely to ensure the product specificity of individual reactions. After qPCR, reactions with multiple products were not taken into account for further analysis. Data were analyzed with Microsoft Excel 2010, applying the $\Delta\Delta C_t$ method. Expression rates were normalized for variation

against the reference gene, a thioredoxin (Pp3c19_1800V3.1), which showed the smallest deviation among a broad range of microarray experiments, including juvenile gametophores, adult gametophores, and green and brown sporophytes (Hiss *et al.*, 2014). For the selection of candidate genes and primer sequences, see Table S2.

### ID conversion

Throughout the text the most recent COSMOSS 3.3 gene identifiers (CGIs) are used. Table S4 shows the corresponding CGIs for the v1.6 annotation and Phypa-IDs (v1.2 annotation) for all genes discussed here.

### Gene ontology (GO) analyses and visualization

The GO bias analyses used Fisher's exact test to calculate *P* values, as described previously (Widiez *et al.*, 2014). Multiple testing-corrected (Benjamini and Hochberg, 1995) *q* values were calculated in R with the function p.adjust (R Development Core Team, 2008). Word-cloud visualizations were created using the online tool WORDLE (http://www.wordle.net). The size of the word is proportional to the −log10(*q* value), and over-represented GO terms were colored dark green if $q \leq 0.0001$ and light green if $q > 0.0001$. Under-represented GO terms were colored dark red if $q \leq 0.0001$ and light red if $q > 0.0001$.

### CIRCOS plots

For the integrative visualization of the individual genomic features one karyotype ideogram was created and tracks were plotted with CIRCOS 0.67-6 (Krzywinski *et al.*, 2009). Chromosomes were split into smaller windows (window size, 500 kbp; window overlaps/jumps, 400 kbp) using values window averages (VWAs), normalized by scaling between a range of 0 and 1 per chromosome using the following equation:

$$\textit{normalized window averagechr vwaichr} = \textit{vwaichr} - \textit{vwachrminvwachrmax} - \textit{vwachrmin}.$$

## ACCESSION NUMBERS

NimbleGen microarray data for juvenile and adult gametophores, and green and brown sporophytes, are available at ArrayExpress: E-MTAB-4630 (http://www.ebi.ac.uk/arrayexpress). Published data for brown sporophytes (Becker lab): E-MTAB-3069. Combimatrix microarray data for brown sporophytes were deposited at ArrayExpress: E-MTAB-916.

Reute gDNA raw data have been made available via the Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra; SRP068341).

## ACKNOWLEDGEMENTS

The JGI Plant Gene Atlas project conducted by the US Department of Energy Joint Genome Institute was supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231. Full Gene Atlas data sets are available at http://phytozome.jgi.doe.gov. This work was supported by the German Federal Ministry of Education and Research (Freiburg Initiative for Systems Biology, 0313921 to SAR). We are grateful to Andrew C. Cuming for language editing. We thank Jörg Becker for access to the sporophyte microarray data prior to publication, and Eva Bieler, Swiss Nanoscience Institute (SNI), Nano Imaging, University of Basel, for help with taking the Cryo-SEM pictures, as well as Faezeh Donges and Marco Göttig for technical assistance, the Deutscher Wetterdienst (DWD) for providing climate data, and Juliet C. Coates and Jörg Becker for comments on the manuscript. The authors declare no conflicts of interest.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

**Figure S1.** Site and weather data for the Reute collection site.
**Figure S2.** Weather data of the Gransden collection site.
**Figure S3.** Bar chart of microarray expression values for the GRAS family protein Pp3c2_20930V3.1.
**Figure S4.** Phylogenetic tree of CHS genes modified after Wolf *et al.* (2010).
**Figure S5.** Word cloud of gene ontology terms (biological process) of the 15 genes confidently expressed at a higher level in mature sporophytes.
**Figure S6.** Box plot of microarray experiments from Reute developmental stages.
**Figure S7.** Hierarchical clustering of microarray experiments from Reute developmental stages.
**Table S1.** Published microarray and RNA-seq data sets for *Physcomitrella patens*.
**Table S2.** Average weather data from the Gransden and Reute sites.
**Table S3.** Primer sequences used for quantitative real-time PCR.
**Table S4.** Gene IDs for the genes discussed here.
**Table S5.** Reute genes that contain an SNP or an Indel leading to a premature stop codon.
**Table S6.** SNP peaks from Reute (as compared with Gransden).
**Table S7.** Over-represented gene ontology terms of genes found in SNP peaks.
**Table S8.** Gene IDs and their annotation for genes found in the chromosome-19 SNP peak.
**Table S9.** List of genes that show a $K_a/K_s$ ratio >2 between Reute and Gransden.
**Appendix S1.** List of SNPs detected by the *P. patens* bead array.
**Appendix S2.** Differentially expressed genes between Gransden and Reute adult gametophores, between Reute adult gametophores and green sporophytes, and between Reute green sporophytes and brown sporophytes.

## REFERENCES

**Ashton, N.W. and Raju, M.V.S.** (2000) The distribution of gametangia on gametophores of Physcomitrella (Aphanoregma) patens in culture. *J. Bryol.* **22**, 9–12.
**Barker, E.I. and Ashton, N.W.** (2013) A parsimonious model of lineage-specific expansion of MADS-box genes in Physcomitrella patens. *Plant Cell Rep.* **32**, 1161–1177.
**Beike, A.K., Horst, N.A. and Rensing, S.A.** (2010) Axenic bryophyte *in vitro* cultivation. *Endocyt Cell Res.* **20**, 102–108.
**Beike, A.K., von Stackelberg, M., Schallenberg-Rudinger, M., Hanke, S.T., Follo, M., Quandt, D., McDaniel, S.F., Reski, R., Tan, B.C. and Rensing, S.A.** (2014) Molecular evidence for convergent evolution and allopolyploid speciation within the Physcomitrium-Physcomitrella species complex. *BMC Evol. Biol.* **14**, 158.
**Beike, A.K., Lang, D., Zimmer, A.D., Wust, F., Trautmann, D., Wiedemann, G., Beyer, P., Decker, E.L. and Reski, R.** (2015) Insights from the cold transcriptome of Physcomitrella patens: global specialization pattern of conserved transcriptional regulators and identification of orphan genes involved in cold acclimation. *New Phytol.* **205**, 869–881.
**Belling, J.** (1921) On counting chromosomes in pollen-mother cells. *Am. Nat.* **55**, 573–574.
**Benjamini, Y. and Hochberg, Y.** (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* **57**, 289–300.

**Bolger, A.M., Lohse, M. and Usadel, B.** (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
**Bolle, C.** (2004) The role of GRAS proteins in plant signal transduction and development. *Planta*, **218**, 683–692.
**Brown, R.C. and Lemmon, B.E.** (2011) Spores before sporophytes: hypothesizing the origin of sporogenesis at the algal-plant transition. *New Phytol.* **2011**, 1469–8137.
**Bryan, V.S.** (1957) Cytotaxonomic studies in the Ephemeraceae and Funariaceae. *The Bryologist*, **60**, 103–126.
**Busch, H., Boerries, M., Bao, J., Hanke, S.T., Hiss, M., Tiko, T. and Rensing, S.A.** (2013) Network theory inspired analysis of time-resolved expression data reveals key players guiding *P. patens* stem cell development. *PLoS ONE*, **8**, e60494.
**Cao, J., Schneeberger, K., Ossowski, S. *et al.*** (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **43**, 956–963.
**Charlot, F., Chelysheva, L., Kamisugi, Y. *et al.*** (2014) RAD51B plays an essential role during somatic and meiotic recombination in Physcomitrella. *Nucleic Acids Res.* **42**, 11965–11978.
**Cingolani, P., Platts, A., le Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M.** (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.
**Colpitts, C.C., Kim, S.S., Posehn, S.E., Jepson, C., Kim, S.Y., Wiedemann, G., Reski, R., Wee, A.G., Douglas, C.J. and Suh, D.Y.** (2011) PpASCL, a moss ortholog of anther-specific chalcone synthase-like enzymes, is a hydroxyalkylpyrone synthase involved in an evolutionarily conserved sporopollenin biosynthesis pathway. *New Phytol.* **192**, 855–868.
**Cuming, A.C., Cho, S.H., Kamisugi, Y., Graham, H. and Quatrano, R.S.** (2007) Microarray analysis of transcriptional responses to abscisic acid and osmotic, salt, and drought stress in the moss, Physcomitrella patens. *New Phytol.* **176**, 275–287.
**Daku, R.M., Rabbi, F., Buttigieg, J., Coulson, I.M., Horne, D., Martens, G., Ashton, N.W. and Suh, D.Y.** (2016) PpASCL, the physcomitrella patens anther-specific chalcone synthase-like enzyme implicated in sporopollenin biosynthesis, is needed for integrity of the moss spore wall and spore viability. *PLoS ONE*, **11**, e0146817.
**Dellaporta, S., Wood, J. and Hicks, J.** (1983) A plant DNA minipreparation: version II. *Plant Mol. Biol. Rep.* **1**, 19–21.
**Engel, P.P.** (1968) The induction of biochemical and morphological mutants in the moss *Physcomitrella patens*. *Am. J. Bot.* **55**, 438–446.
**Frank, W., Decker, E.L. and Reski, R.** (2005) Molecular tools to study *Physcomitrella patens*. *Plant Biol. (Stuttg)* **7**, 220–227.
**Frey, W., Stech, M. and Fischer, E.** (2009) *Syllabus of Plant Families – Part 3 Bryophytes and Seedless Vascular Plants Berlin*. Stuttgart: Borntraeger.
**Glime, J.M.** (2007) Bryophyte ecology. In *Volume 1. Physiological Ecology* (Glime, J.M. ed) Ebook sponsored by Michigan Technological University and the International Association of Bryologists. Available at http://www.bryoecol.mtu.edu/.
**Gramzow, L. and Theissen, G.** (2010) A hitchhiker's guide to the MADS world of plants. *Genome Biol.* **11**, 214.
**Hiss, M., Laule, O., Meskauskiene, R.M. *et al.*** (2014) Large-scale gene expression profiling data for the model moss Physcomitrella patens aid understanding of developmental progression, culture and stress conditions. *Plant J.* **79**, 530–539.
**Hoffmann, G.R.** (1970) Spore viability in *Funaria hygrometrica*. *Bryologist*, **73**, 634–635.
**Hohe, A., Rensing, S.A., Mildner, M., Lang, D. and Reski, R.** (2002) Day length and temperature strongly influence sexual reproduction and expression of a novel MADS-Box gene in the moss *Physcomitrella patens*. *Plant Biol.* **4**, 762–762.
**Horst, N.A., Katz, A., Pereman, I., Decker, E.L., Ohad, N. and Reski, R.** (2016) A single homeobox gene triggers phase transition, embryogenesis and asexual reproduction. *Nat. Plants*, **2**, 15209.
**Hutter, S., Vilella, A.J. and Rozas, J.** (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, **7**, 409.
**Kamisugi, Y. and Cuming, A.C.** (2005) The evolution of the abscisic acid-response in land plants: comparative analysis of group 1 LEA gene expression in moss and cereals. *Plant Mol. Biol.* **59**, 723–737.
**Kamisugi, Y., von Stackelberg, M., Lang, D., Care, M., Reski, R., Rensing, S.A. and Cuming, A.C.** (2008) A sequence-anchored genetic linkage map for the moss, Physcomitrella patens. *Plant J.* **56**, 855–866.
**Khandelwal, A., Cho, S.H., Marella, H., Sakata, Y., Perroud, P.F., Pan, A. and Quatrano, R.S.** (2010) Role of ABA and ABI3 in desiccation tolerance. *Science*, **327**, 546.
**Knop, W.** (1868) *Der Kreislauf des Stoffs: Lehrbuch der Agricultur-Chemie*. Leipzig: H. Haessel.
**Koduri, P.K.H., Gordon, G., Barker, E., Colpitts, C., Ashton, N. and Suh, D.-Y.** (2010) Genome-wide analysis of the chalcone synthase superfamily genes of Physcomitrella patens. *Plant Mol. Biol.* **72**, 247–263.
**Komatsu, K., Nishikawa, Y., Ohtsuka, T., Taji, T., Quatrano, R.S., Tanaka, S. and Sakata, Y.** (2009) Functional analyses of the ABI1-related protein phosphatase type 2C reveal evolutionarily conserved regulation of abscisic acid signaling between Arabidopsis and the moss Physcomitrella patens. *Plant Mol. Biol.* **6**, 6.
**Koressaar, T. and Remm, M.** (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics*, **23**, 1289–1291.
**Krupa, J.** (1967) Studies on the physiology of germination of spores of *Funaria hygrometrica*. III. The influence of monochromatic light on the germination of the spores. *Acta Soc. Bot. Polon.* **36**, 57–65.
**Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A.** (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.
**Landberg, K., Pederson, E.R., Viaene, T., Bozorg, B., Friml, J., Jonsson, H., Thelander, M. and Sundberg, E.** (2013) The MOSS Physcomitrella patens reproductive organ development is highly organized, affected by the two SHI/STY genes and by the level of active auxin in the SHI/STY expression domain. *Plant Physiol.* **162**, 1406–1419.
**Lang, D., Zimmer, A.D., Rensing, S.A. and Reski, R.** (2008) Exploring plant biodiversity: the Physcomitrella genome and beyond. *Trends Plant Sci.* **13**, 542–549.
**Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.** and **1000 Genome Project Data Processing Subgroup**. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
**Li, S., Zhao, Y., Zhao, Z., Wu, X., Sun, L., Liu, Q. and Wu, Y.** (2016) Crystal structure of the GRAS domain of SCARECROW-LIKE 7 in Oryza sativa. *Plant Cell*, **28**, 1025–1034.
**Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y.** (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517.
**McDaniel, S.F., von Stackelberg, M., Richardt, S., Quatrano, R.S., Reski, R. and Rensing, S.A.** (2010) The speciation history of the Physcomitrium-Physcomitrella species complex. *Evolution*, **64**, 217–231.
**McKenna, A., Hanna, M., Banks, E. *et al.*** (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.
**Medina, R., Liu, Y., Li-Song, W., Shuiliang, G., Hylander, K. and Goffinet, B.** (2015) DNA based revised geographic circumscription of species of Physcomitrella sl (Funariaceae): *P. patens* new to East Asia and *P. magdalenae* new to East Africa. *The Bryologist*, **1**, 22–31.
**Morohashi, K., Minami, M., Takase, H., Hotta, Y. and Hiratsuka, K.** (2003) Isolation and characterization of a novel GRAS gene that regulates meiosis-associated gene expression. *J. Biol. Chem.* **278**, 20865–20873.
**Mosquna, A., Katz, A., Decker, E.L., Rensing, S.A., Reski, R. and Ohad, N.** (2009) Regulation of stem cell maintenance by the Polycomb protein FIE has been conserved during land plant evolution. *Development*, **136**, 2433–2444.
**Nakosteen, P.C. and Hughes, K.W.** (1978) Sexual Life cycle of three species of Funariaceae in culture. *Bryologist*, **81**, 307–314.
**O'Donoghue, M.T., Chater, C., Wallace, S., Gray, J.E., Beerling, D.J. and Fleming, A.J.** (2013) Genome-wide transcriptomic analysis of the sporophyte of the moss Physcomitrella patens. *J. Exp. Bot.* **64**, 3567–3581.
**Okano, Y., Aono, N., Hiwatashi, Y., Murata, T., Nishiyama, T., Ishikawa, T., Kubo, M. and Hasebe, M.** (2009) A polycomb repressive complex 2 gene regulates apogamy and gives evolutionary insights into early land plant evolution. *Proc. Natl Acad. Sci. USA*, **106**, 16321–16326.
**Ortiz-Ramirez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., Feijo, J.A. and Becker, J.D.** (2015) A transcriptome atlas of

Physcomitrella patens provides insights into the evolution and development of land plants. *Mol. Plant*, **9**, 205–220.

**Perroud, P.F., Cove, D.J., Quatrano, R.S. and McDaniel, S.F.** (2011) An experimental method to facilitate the identification of hybrid sporophytes in the moss Physcomitrella patens using fluorescent tagged lines. *New Phytol.* **2**, 1469–8137.

**Possart, A. and Hiltbrunner, A.** (2013) An evolutionarily conserved signaling mechanism mediates far-red light responses in land plants. *Plant Cell*, **25**, 102–114.

**Prigge, M.J. and Bezanilla, M.** (2010) Evolutionary crossroads in developmental biology: *Physcomitrella patens*. *Development*, **137**, 3535–3543.

**Quatrano, R.S., McDaniel, S.F., Khandelwal, A., Perroud, P.F. and Cove, D.J.** (2007) *Physcomitrella patens*: mosses enter the genomic age. *Curr. Opin. Plant Biol.* **10**, 182–189.

**R Development Core Team** (2008) *R: A Language and Environment for Statistical Computing*. Available at http://www.R-project.org/.

**Rensing, S.A., Ick, J., Fawcett, J.A., Lang, D., Zimmer, A., Van de Peer, Y. and Reski, R.** (2007) An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol. Biol.* **7**, 130.

**Rensing, S.A., Lang, D., Zimmer, A.D. *et al.*** (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.

**Rensing, S.A., Beike, A.K. and Lang, D.** (2012) Evolutionary importance of generative polyploidy for genome evolution of haploid-dominant land plants. In *Plant Genome Diversity* (Greilhuber, J., Wendel, J.F., Leitch, I.J. and Doležel, J. eds). Vienna, New York: Springer, pp. 295–305

**Reski, R. and Cove, D.J.** (2004) Quick guide: *Physcomitrella patens*. *Curr. Biol.* **14**, R261–R262.

**Sakakibara, K., Nishiyama, T., Deguchi, H. and Hasebe, M.** (2008) Class 1 KNOX genes are not involved in shoot development in the moss Physcomitrella patens but do function in sporophyte development. *Evol. Dev.* **10**, 555–566.

**Sakakibara, K., Ando, S., Yip, H.K., Tamada, Y., Hiwatashi, Y., Murata, T., Deguchi, H., Hasebe, M. and Bowman, J.L.** (2013) KNOX2 genes regulate the haploid-to-diploid morphological transition in land plants. *Science*, **339**, 1067–1070.

**Schruth, D.** (2012) *Caroline: A Collection of Database, Data Structure, Visualization, and Utility Functions for R, R package version 0.7. 4*. Available at https://rdrr.io/cran/caroline/.

**von Stackelberg, M., Rensing, S.A. and Reski, R.** (2006) Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. *BMC Plant Biol.* **6**, 9.

**Stevenson, S.R., Kamisugi, Y., Trinh, C.H. *et al.*** (2016) Genetic analysis of Physcomitrella patens identifies ABSCISIC ACID NON-RESPONSIVE (ANR), a regulator of ABA responses unique to basal land plants and required for desiccation tolerance. *Plant Cell.* **28**, 1310–1327.

**Sun, X., Xue, B., Jones, W.T., Rikkerink, E., Dunker, A.K. and Uversky, V.N.** (2011) A functionally required unfoldome from the plant kingdom: intrinsically disordered N-terminal domains of GRAS proteins are involved in molecular recognition during plant development. *Plant Mol. Biol.* **77**, 205–223.

**Szovenyi, P., Devos, N., Weston, D.J., Yang, X., Hock, Z., Shaw, J.A., Shimizu, K.K., McDaniel, S.F. and Wagner, A.** (2014) Efficient purging of deleterious mutations in plants with haploid selfing. *Genome Biol. Evol.* **6**, 1238–1252.

**Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G.** (2012) Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115.

**Vesty, E.F., Saidi, Y., Moody, L.A. *et al.*** (2016) The decision to germinate is regulated by divergent molecular networks in spores and seeds. *New Phytol.* **211**, 952–966.

**Wang, X.Y. and Paterson, A.H.** (2011) Gene conversion in angiosperm genomes with an emphasis on genes duplicated by polyploidization. *Genes (Basel)*, **2**, 1–20.

**Wang, D., Zhang, Y., Zhang, Z., Zhu, J. and Yu, J.** (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics*, **8**, 77–80.

**Wang, X., Tang, H. and Paterson, A.H.** (2011) Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. *Plant Cell*, **23**, 27–37.

**Widiez, T., Symeonidi, A., Luo, C., Lam, E., Lawton, M. and Rensing, S.A.** (2014) The chromatin landscape of the moss Physcomitrella patens and its dynamics during development and drought stress. *Plant J.* **79**, 67–81.

**Wolf, L., Rizzini, L., Stracke, R., Ulm, R. and Rensing, S.A.** (2010) The Molecular and Physiological Responses of Physcomitrella patens to Ultraviolet-B Radiation. *Plant Physiol.* **153**, 1123–1134.

**Wu, T.D. and Nacu, S.** (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

**Xiao, L., Wang, H., Wan, P., Kuang, T. and He, Y.** (2011) Genome-wide transcriptome analysis of gametophyte development in Physcomitrella patens. *BMC Plant Biol.* **11**, 177.

**Xiao, L., Zhang, L., Yang, G., Zhu, H. and He, Y.** (2012) Transcriptome of protoplasts reprogrammed into stem cells in Physcomitrella patens. *PLoS ONE*, **7**, e35961.

**Yaari, R., Noy-Malka, C., Wiedemann, G., Auerbach Gershovitz, N., Reski, R., Katz, A. and Ohad, N.** (2015) DNA METHYLTRANSFERASE 1 is involved in (m)CG and (m)CCG DNA methylation and is essential for sporophyte development in Physcomitrella patens. *Plant Mol. Biol.* **88**, 387–400.

**Zhao, S., Fung-Leung, W.P., Bittner, A., Ngo, K. and Liu, X.** (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, **9**, e78644.

**Zimmer, A.D., Lang, D., Buchta, K., Rombauts, S., Nishiyama, T., Hasebe, M., Van de Peer, Y., Rensing, S.A. and Reski, R.** (2013) Reannotation and extended community resources for the genome of the non-seed plant Physcomitrella patens provide insights into the evolution of plant gene structures and functions. *BMC Genomics*, **14**, 498.

## 6.3 Publication 3

In this recent publication I added the Kaskaskia ecotype to the SNP analysis that shows a higher SNP density than Reute, but is not as divergent from Gransden as the Villersexel K3 ecotype. For all three ecotypes we also investigated the SNPs across the pseudochromosomal genome sequences and could find hot spots with higher SNP densities. One of these hot spots is present in all three investigated ecotypes and the hot spot region contains genes that are part of the sterol catabolism and genes that are annotated as part of the chloroplast light sensing/movement.

Detailed sequence information of different ecotypes is used e. g. in *Arabidopsis thaliana* to identify how genetic variation contributes to adaptation to diverse environments (Cao *et al.*, 2011). Including more accessions into the sequence analysis can help detect regional differences and to predict distribution patterns. From our SNP analysis it seems that closer regional proximity cannot be inferred from the genotype since the Villersexel K3 ecotype is geographically equally far from Gransden as is the Reute accession but shows a much higher genetic diversity from the Gransden ecotype. A possible explanation for this pattern is that spores from *P. patens* may be distributed rather by birds than by other means. Therefore a close genetic distance can also be found in geographically distant accessions.

the plant journal

SEB Society for Experimental Biology

*The Plant Journal* (2018) **93**, 515–533 doi: 10.1111/tpj.13801

# The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution

**Daniel Lang[1,2,#], Kristian K. Ullrich[3,#,†], Florent Murat[4], Jörg Fuchs[5], Jerry Jenkins[6], Fabian B. Haas[3], Mathieu Piednoel[7], Heidrun Gundlach[2], Michiel Van Bel[8,9], Rabea Meyberg[3], Cristina Vives[10], Jordi Morata[10], Aikaterini Symeonidi[3,‡], Manuel Hiss[3], Wellington Muchero[11], Yasuko Kamisugi[12], Omar Saleh[1,§], Guillaume Blanc[13], Eva L. Decker[1], Nico van Gessel[1], Jane Grimwood[6,14], Richard D. Hayes[14], Sean W. Graham[15], Lee E. Gunter[11], Stuart F. McDaniel[16], Sebastian N.W. Hoernstein[1], Anders Larsson[17], Fay-Wei Li[18], Pierre-François Perroud[3], Jeremy Phillips[14], Priya Ranjan[11], Daniel S. Rokshar[14,19], Carl J. Rothfels[20], Lucas Schneider[3,¶], Shengqiang Shu[14], Dennis W. Stevenson[21], Fritz Thümmler[22], Michael Tillich[23], Juan C. Villarreal Aguilar[24], Thomas Widiez[25,26,**], Gane Ka-Shu Wong[27,28,29], Ann Wymore[11], Yong Zhang[30], Andreas D. Zimmer[1,††], Ralph S. Quatrano[31], Klaus F.X. Mayer[2,32], David Goodstein[14], Josep M. Casacuberta[10], Klaas Vandepoele[8,9], Ralf Reski[1,33], Andrew C. Cuming[12], Gerald A. Tuskan[11], Florian Maumus[34], Jérome Salse[4], Jeremy Schmutz[6,14] and Stefan A. Rensing[3,33,*]**

[1]*Plant Biotechnology, Faculty of Biology, University of Freiburg, Schaenzlestr. 1, 79104, Freiburg, Germany,*
[2]*Plant Genome and Systems Biology, Helmholtz Center Munich, 85764, Neuherberg, Germany,*
[3]*Plant Cell Biology, Faculty of Biology, University of Marburg, Marburg, Germany,*
[4]*INRA, UMR 1095 Genetics, Diversity and Ecophysiology of Cereals (GDEC), 5 Chemin de Beaulieu, 63100, Clermont-Ferrand, France,*
[5]*Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, OT Gatersleben, D-06466, Stadt Seeland, Germany,*
[6]*HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA,*
[7]*Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné Weg 10, D-50829, Cologne, Germany,*
[8]*VIB Center for Plant Systems Biology, Technologiepark 927, 9052 Ghent, Belgium,*
[9]*Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052, Gent, Belgium,*
[10]*Center for Research in Agricultural Genomics, CRAG (CSIC-IRTA-UAB-UB), Campus UAB, Bellaterra, Cerdanyola del Vallès, 08193, Barcelona, Spain,*
[11]*Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA,*
[12]*Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK,*
[13]*Structural and Genomic Information Laboratory (IGS), Aix-Marseille Université, CNRS, UMR 7256 (IMM FR 3479), Marseille, France,*
[14]*DOE Joint Genome Institute, Walnut Creek, CA 94598, USA,*
[15]*Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada,*
[16]*Department of Biology, University of Florida, Gainesville, FL 32611, USA,*
[17]*Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden,*
[18]*Boyce Thompson Institute, Ithaca, NY 14853, USA,*
[19]*Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA,*
[20]*University Herbarium and Department of Integrative Biology, University of California, Berkeley, CA 94720-2465, USA,*
[21]*New York Botanical Garden, Bronx, NY 10458, USA,*
[22]*Vertis Biotechnologie AG, Lise-Meitner-Str. 30, 85354, Freising, Germany,*
[23]*Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476, Potsdam-Golm, Germany,*
[24]*Department of Biology, Université* Laval, Québec G1V 0A6, Canada,
[25]*Department of Plant Biology, University of Geneva, Sciences III, Geneva 4 CH-1211, Switzerland,*
[26]*Department of Plant Biology &* Pathology Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA,
[27]*Department of Biological Sciences, University of Alberta, Edmonton, AB, T6G 2E9, Canada,*
[28]*Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada,*
[29]*BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China,*
[30]*Shenzhen Huahan Gene Life Technology Co. Ltd, Shenzhen, China,*
[31]*Department of Biology, Washington University, St. Louis, MO, USA,*
[32]*WZW, Technical University Munich, Munich, Germany,*
[33]*BIOSS Centre for Biological Signalling Studies, University of Freiburg, Schaenzlestr. 18, 79104, Freiburg, Germany,*
[34]*URGI, INRA, Université* Paris-Saclay, 78026, Versailles, France,

Received 11 October 2017; revised 20 November 2017; accepted 24 November 2017; published online 13 December 2017.
*For correspondence (e-mail stefan.rensing@biologie.uni-marburg.de).
#These authors contributed equally to this work.
† Present address: Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306, Ploen, Germany.
‡ Present address: Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain.
§ Present address: Plant Molecular Cell Biology, Humboldt-University of Berlin, 10115, Berlin, Germany.
¶ Present address: Institute for Transfusion Medicine and Immunohematology, Johann-Wolfgang-Goethe University and German Red Cross Blood Service, Sandhofstraße 1, 60528, Frankfurt am Main, Germany.
** Present address: Laboratoire Reproduction et D_eveloppement des Plantes, Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, INRA, F-69342, Lyon, France.
†† Present address: Faculty of Medicine, Institute for Human Genetics, Medical Center – University of Freiburg, Freiburg, Germany.

## SUMMARY

**The draft genome of the moss model, *Physcomitrella patens*, comprised approximately 2000 unordered scaffolds. In order to enable analyses of genome structure and evolution we generated a chromosome-scale genome assembly using genetic linkage as well as (end) sequencing of long DNA fragments. We find that 57% of the genome comprises transposable elements (TEs), some of which may be actively transposing during the life cycle. Unlike in flowering plant genomes, gene- and TE-rich regions show an overall even distribution along the chromosomes. However, the chromosomes are mono-centric with peaks of a class of Copia elements potentially coinciding with centromeres. Gene body methylation is evident in 5.7% of the protein-coding genes, typically coinciding with low GC and low expression. Some giant virus insertions are transcriptionally active and might protect gametes from viral infection *via* siRNA mediated silencing. Structure-based detection methods show that the genome evolved *via* two rounds of whole genome duplications (WGDs), apparently common in mosses but not in liverworts and hornworts. Several hundred genes are present in colinear regions conserved since the last common ancestor of plants. These syntenic regions are enriched for functions related to plant-specific cell growth and tissue organization. The *P. patens* genome lacks the TE-rich pericentromeric and gene-rich distal regions typical for most flowering plant genomes. More non-seed plant genomes are needed to unravel how plant genomes evolve, and to understand whether the *P. patens* genome structure is typical for mosses or bryophytes.**

**Keywords: evolution, genome, chromosome, plant, moss, methylation, duplication, synteny, *Physcomitrella patens*.**

## INTRODUCTION

The original genome sequencing of the model moss *Physcomitrella patens* (Hedw.) Bruch & Schimp. (Funariaceae) reflected its informative phylogenetic position: a very early divergence from the evolutionary path that eventually led to the flowering plants soon after the first plants conquered land *ca.* 500 Ma ago (Lang *et al.*, 2010). Previous comparisons of the moss genome with those of flowering plants and green algae provided many insights into land plant evolution (Rensing *et al.*, 2008), detailing for example the evolution of abiotic stress responses and phytohormone signaling. Subsequent comparative functional genomic analyses, making use of the ability of *P. patens* for 'reverse genetics' by gene targeting, addressed questions of how gene functions evolved to enable the increasing developmental and anatomical complexity that characterizes the dominant forms of plant life on the planet (e.g. Horst *et al.*, 2016; Sakakibara *et al.*, 2013). The initial draft sequence encompassed close to 2000 unordered scaffolds, significantly limiting analyses of chromosomal structure and evolution, or of the conservation of gene order during land plant evolution. We now present a new assembly accurately representing the chromosomal architecture (pseudochromosomes). Much-increased acquisition of transcriptomic evidence has substantially improved the quality of gene annotation, and acquisition of high-density DNA methylation and histone mark data combined with a detailed analysis of transposable elements (TEs) explain the size and architecture of the moss genome. This study provides unprecedented insights into the genome of a haploid-dominant land plant, such as the peculiar structure and evolution of moss chromosomes, and demonstrates syntenic conservation of important plant genes throughout 500 Ma of evolution.

## RESULTS AND DISCUSSION

### The moss V3 genome: assembly and annotation

The original genome sequence (V1.2) of *Physcomitrella patens* (strain Gransden 2004) comprised 1995 sequence

scaffolds (Rensing *et al.*, 2008; Zimmer *et al.*, 2013). Here, we integrated the previous sequence data with a high-density genetic linkage map based on 3712 SNP segregating loci in a cross between the 'Gransden 2004' (Gransden) laboratory strain and the genetically divergent 'Villersexel K3' (Villersexel) accession (Kamisugi *et al.*, 2008). The resulting assembly was further improved using novel BAC/fosmid paired end sequence data (cf. Appendix S1, Supplementary Material I for details; see section Availability of gene models and additional data for novel data associated with this study). We screened the subsequent integrated assembly for sequence contamination, producing a pseudomolecule release covering 27 nuclear chromosomes with a total genetic linkage distance of 5502.6–5503.1 centiMorgans (cM). The 27 chromosomal pseudomolecules include 462.3 Mbp of sequence, supplemented by 351 unplaced scaffolds representing 4.9 Mbp (1%) of unintegrated sequence, totaling 90% of the 518 Mbp estimated by flow cytometry (Schween *et al.*, 2003). The reads partitioned as mitochondrial and plastidal were assembled *de novo*, yielding an improved assembly and annotation of both organellar genomes (correcting e.g. the N-terminal sequence of the plastidal RuBisCO). Structural annotation used substantial new transcript evidence (File S3). For parameter optimization it relied on a manually curated reference gene set (Zimmer *et al.*, 2013), yielding gene annotation version 3.1. Of 35 307 predicted protein-coding genes, 27 511 (78%) could be functionally annotated (cf. Appendix S1, Supplementary Material II and File S1), i.e. encode known domains and/or encode homologs of proteins in other species. In total, 20 274 (57%) genes are expressed based on RNA-seq evidence of typical developmental stages covered by the JGI gene atlas project (http://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/); the remaining genes might be expressed in as yet unrepresented stages such as mature spores or male gametes. We found 13 160 genes to be expressed in the juvenile gametophyte (Figure 1), the filamentous protonemata, 12 714 in the adult gametophyte, the leafy gametophores, and 14 309 in the diploid sporophytes developing from the zygote (overlap: 10 388 genes expressed in all three developmental stages).

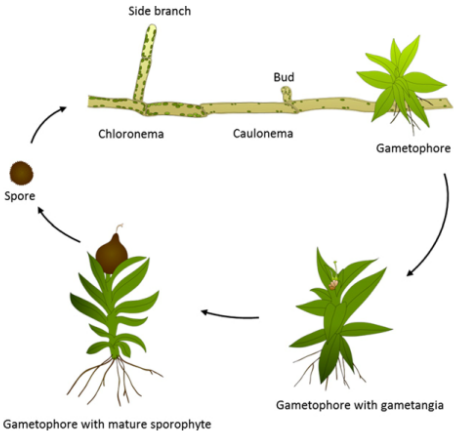



**Figure 1.** The *P. patens* life cycle.
Germination of haploid spores yields the juvenile gametophytic generation, the protonema. Protonema grows two-dimensional by apical (tip) growth and side branching. Protonemata consist of chloroplast-rich chloronema cells, and longer, thinner caulonema cells featuring less chloroplasts and oblique cross walls. Three-faced buds featuring single apical stem cells emerge from side branches (Harrison *et al.*, 2009) to form the adult gametophytic phase, the leafy gametophores. Gametophores comprise basal, multicellular rhizoids for nutrient supply, as well as non-vascular leaves (phyllids). Gametangia (female archegonia and male antheridia) develop on the gametophores. Upon fertilization of the egg cell by motile spermatozoids the diploid zygote forms and subsequently performs embryogenesis. Spore mother cells in the diploid sporophyte undergo meiosis to form spores.

### Unusual genome structure

*Transposon content and activity.* *De novo* analyses of repeated sequences revealed that the genome is highly repetitive, with 57% of the assembly comprising TEs, tandem repeats, unclassified repeats, and segments of host genes (cf. Appendix S1, Supplementary Material III and Table S13). The vast majority of TEs are long terminal repeat (LTR) retrotransposons (RT), strongly dominated by Gypsy-type elements that contribute almost 48%, with Copia-type elements much less abundant (3.5%). The estimated relative insertion times of LTR-RTs confirm the limited accumulation of Copia-type elements over a prolonged evolutionary time. By contrast, two peaks of Gypsy-type elements testify to both ancient and recent periods of significant TE activity (Figure S7). Phylogenetic inference revealed the presence of five main LTR-RT groups including three Gypsy-type (RLG1-3) and two Copia-type elements (RLC4-5; Figure S8). Applying a molecular clock based on sequence divergence to the full length, intact LTR-RTs indicates that the latest (<1 Ma) activity of Gypsy-type elements was mostly contributed by RLG1-3 elements, preceded by the amassing of RLG2 and RLC5 copies (around 4–6 Ma, Figures S7 and S36). RLG1 thus comprises the youngest and most abundant group among intact LTR-RTs. In line with these results, analysis of TE insertion polymorphisms between Gransden and Villersexel showed that RLG1 elements are highly polymorphic, accounting for most of the detected insertion variants (Figure S9). Since we detect such insertions in both accessions, the decades long *in vitro* culture of Gransden is not likely to be the major source of transposon activity. RLG1 elements are expressed in non-stressed protonemata (Figure S6), which is uncommon as transposon expression is usually strongly silenced in

plants and is only detected in very specific tissues such as pollen, in silencing mutants or under stress situations (Martinez and Slotkin, 2012). Moreover, recent data suggest that some stresses that typically induce plant retrotransposons, such as protoplastation, inhibit RLG1 expression (Vives *et al.*, 2016), suggesting that RLG1 may transpose during the *P. patens* life cycle and might play a role in its genome dynamics. The moss germinates from spores that develop into filamentous, tip-growing protonemata (comprising chloroplast-rich chloronemal and fast-growing caulonemal cells; Figure 1). Buds develop from caulonemal cells and grow into gametophores that bear sexual organs (gametangia). Mosses are prone to endopolyploidy (Bainard and Newmaster, 2010) and older *P. patens* caulonema cells endoreduplicate (Schween *et al.*, 2005). Interestingly, endoreduplicated caulonemal cells give rise to somatic sporophytes if PpBELL1 is overexpressed, thus circumventing sexual reproduction (Horst *et al.*, 2016). *De facto* 2n caulonemal cells might constitute a staging ground for (potentially transmitted) somatic changes caused *via* transposon activity.

*Unusual chromatin structure.* The genomes of most flowering plants are typically composed of monocentric chromosomes, whose unique centromeres are surrounded by heterochromatic pericentromeric regions, that are repeat-rich and gene-poor relative to distal (sub-telomeric), euchromatic regions (Lamb *et al.*, 2007; Figure S34). By contrast, the landscape of gene and repeat density along *P. patens* chromosomes is rather homogeneous, we do not detect large repeat-rich regions with relatively low gene density (Figures 2 and 3). At a finer scale, we do detect an alternation of gene-rich and repeat-rich regions all along the chromosomes (Figure S10). Typical plant pericentromeres are more prone to structural variation (e.g. TE insertions and deletions) compared with the remainder of chromosome arms (Li *et al.*, 2014). Yet, analysis of *P. patens* chromosomes failed to identify hotspots of structural variation that could coincide with pericentromeres (Figure S11). It should be noted, however, that the centromeres could be present at least partially in the unassembled parts of the genome. In any case, immuno-labeling of mitotic metaphase chromosomes using a pericentromere-specific antibody demonstrates that they are mono-centric (Figure S5). Unlike in many flowering plant genomes, the *P. patens* chromosomes are characterized by a more uniform distribution of eu- and heterochromatin (Figures 3, S5 and S35), raising questions about the nature and location of centromeres.

*Physcomitrella centromeres seem to coincide with a particular subset of Copia elements.* Plant centromeres typically comprise large arrays of satellite repeats that can be punctuated by some TEs (Wang *et al.*, 2009). However, plotting the density of tandem repeats along the *P. patens* chromosomes did not reveal peaks likely to reflect the position of centromeres (Figure S11). Computational analysis of tandem repeats in a variety of genomes identified candidate centromeric repeats in *P. patens*, although green algae, mosses, and liverworts contain low abundances of these (Melters *et al.*, 2013). Positioning them on the *P. patens* V3 assembly revealed a patchy distribution, not single peaks that could coincide with centromeres as expected for monocentric chromosomes (Figures S5 and S11). By contrast, the low abundance Copia-type elements exhibited unusually discrete density peaks, typically one per assembled chromosome, spanning hundreds of kbp (Figures 2 and S11). Each Copia density peak principally contains RLC5 elements. A similar situation has been described in the green alga *Coccomyxa subellipsoidea* in which a single peak of a LINE-type retrotransposon, the Zepp element, was proposed to be involved in centromeric function (Blanc *et al.*, 2012). The RLC5 density peak regions are generally punctuated by unresolved gaps in the assembly and by fragments of other TEs (Figure S12). Closer examination revealed that they comprise full length LTR-RTs (FL_RLC5) as well as highly similar truncated non-autonomous variants (Tr_RLC5) that lack the integrase (INT) and reverse transcriptase domains (RVT) (Figure S13). Remarkably, all RLC5 clusters appear to be mosaics containing nested insertions of both FL_RLC5 and Tr_RLC5 elements, of which additional copies are rare in the genome. A neutral explanation for the distribution of RLC5 clusters is that their target sequences are present at a single location per chromosome, perhaps caused by a preference for self-insertion. Alternatively, a single cluster combining FL_RLC5 and Tr_RLC5 copies may be necessary for normal chromosome function. In either case, it is possible that RLC5 clusters might be specific components of centromeres in *P. patens*. The dominant RLC5 peak per chromosome, highlighting the putative centromere, is marked by a radius in Figures 2 and 4.

*Alternation of activating and repressing epigenetic marks.* For the V1.2 scaffolds that harbor histone 3 (H3) ChIP-seq evidence (Widiez *et al.*, 2014), 96% can be mapped to the 27 V3 pseudochromosomes (Figure 4); the remaining 4% map to the unassigned V3 scaffolds, underscoring the quality of the assembly. The alternating structure of genes and TE/DNA methylation (purple in Figure 4) over the full length of the chromosomes is mirrored by activating H3 marks (K4me3, K27Ac, K9Ac; green in Figure 4) corresponding to transcribed genic areas, and repressive H3 marks (K27me3, K9me2; red in Figure 4) coinciding with TEs/intergenic areas. This result contrasts sharply with many flowering plant genomes (Figure S34) in which gene-rich chromosome arms display less heterochromatin than pericentromeres. Similar

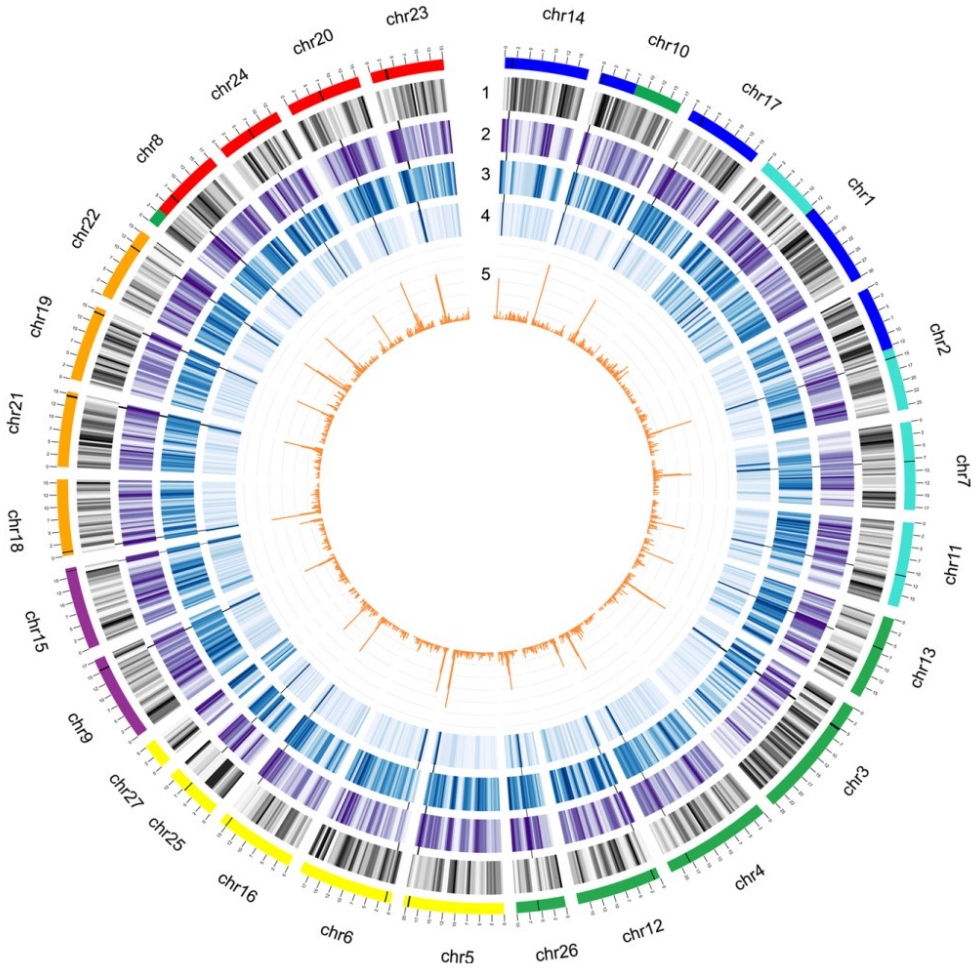



**Figure 2.** Chromosome structure, focus on TEs.
From outer to inner: karyotype bands colored according to ancestral genome blocks as in Figure 5 (scale = Mbp), followed by: (1) gene density (grey, normalized 0,1); (2) repeat density (violet, normalized 0,1); (3) gypsy-type elements (blue, normalized 0,1); (4) Copia-type elements (blue, normalized 0,1); and (5) RLC5 elements (orange, histogram). For each chromosome, a radius marks the dominant RLC5 peak, potentially coinciding with the centromere (see text). All plots are based on a 500 kbp sliding window (400 kbp jump). Chromosomes are arranged according to the ancestral (pre-WGD) seven chromosome karytope (Figure 5).

to flowering plant genomes, TE bodies are generally depleted for histone marks, excepting the silencing mark H3K9me2 that is above background levels in the filamentous protonemata, and at background level in unstressed and stressed leafy gametophores (File S2). The previously described (Widiez *et al.*, 2014) deposition of H3K27me3 at developmental genes that takes place with the switch from protonema to gametophore (Figure 1) can be observed genome-wide (File S2). All TE bodies are methylated in similar fashion, with CG and CHG more abundant than CHH (>80% CG and CHG, >40% CHH; Figures S15 and S25–S28), whereas gene bodies remain barely methylated (Figures S15 and S25–S29). RLC4 has the sharpest boundary pattern (File S2), with almost no methylation outside the TE, followed by RLC5 with more outside-TE methylation, especially CHH. RLG1

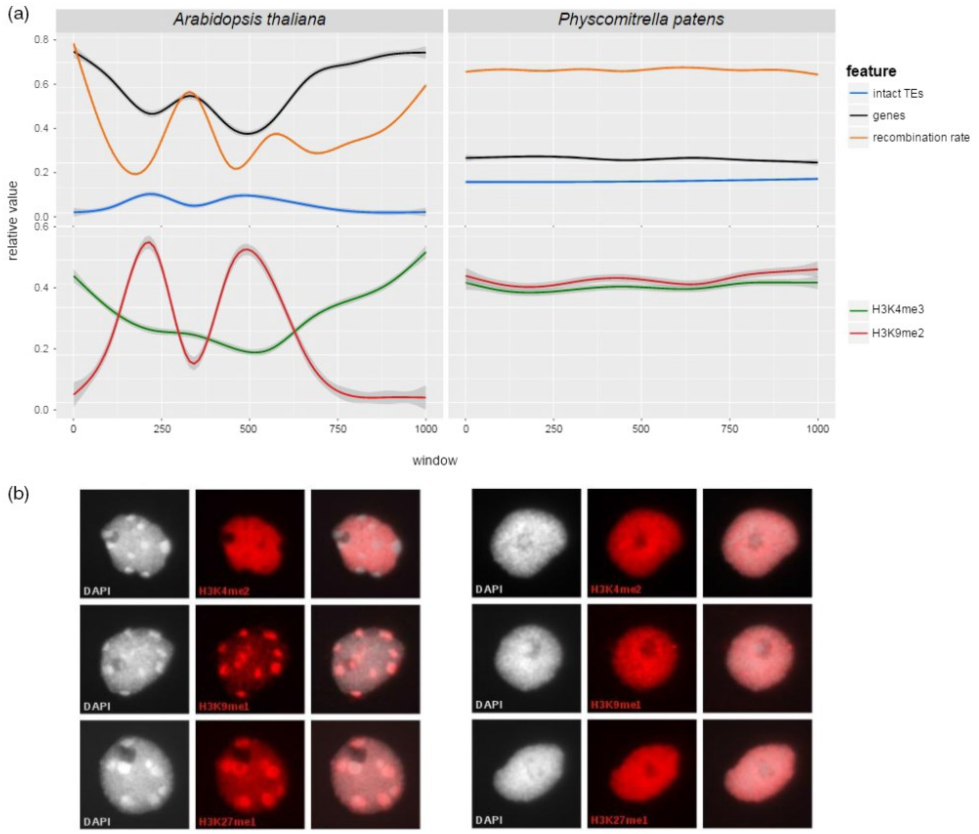



**Figure 3.** Comparative analysis of genome structures.
Comparative data of *Arabidopsis thaliana* (left) and *Physcomitrella patens* (right) reveals the lack of large heterochromatic blocks (b) that is mirrored by even distribution of recombination rate, gene and LTR-RT distribution (a) in the moss.
(a) Averaged topology of genomic features based on 1000 non-overlapping windows per chromosome (averaged over all chromosomes); arbitrary units, 1000 representing the full length of the averaged chromosomes. Upper track: Smoothed chromosomal densities of intact LTRs, protein-coding genes and the normalized mean recombination rate. Lower track: Smoothed density curves of H3K4me3 and H3K9me2 histone modification peak regions.
(b) Immunostaining of typical eu- and heterochromatin-associated histone methylation marks (H3K4me2, H3K9me1 and H3K27me1) on flow-sorted interphase nuclei.

follows in a similar fashion, although the relatively sharp pattern of RLG1 and RLC5 can in part be attributed to the fact that in case of nested insertions no 'outside' TE region is present next to the TE boundary. RLG2 shows a broad pattern of all three contexts, RLG3 shows the broadest pattern with no discernible body peak. As the methylation pattern of the main TE categories differs in how sharply they define the TE proper, TE families might have different impacts on the proximal epigenome.

*Gene body methylation marks low GC genes.* Interestingly, intron-containing genes (Figure S25) show a much sharper methylation contrast between gene body and surrounding DNA, and a more pronounced difference between CHH and the other contexts, than intron-less genes (Figure S26). As the latter genes might in part be retrocopies (Kaessmann, 2010), they might be more prone to silencing and be embedded in more homogeneously methylated areas. Gene-body methylation (GBM) is found in many eukaryotic lineages and is thought to have been present in the last common eukaryotic ancestor (Feng *et al.*, 2010). GBM in flowering plants is characterized by CG methylation of the coding sequence, not extending to transcriptional start and stop (Niederhuth *et al.*, 2016).

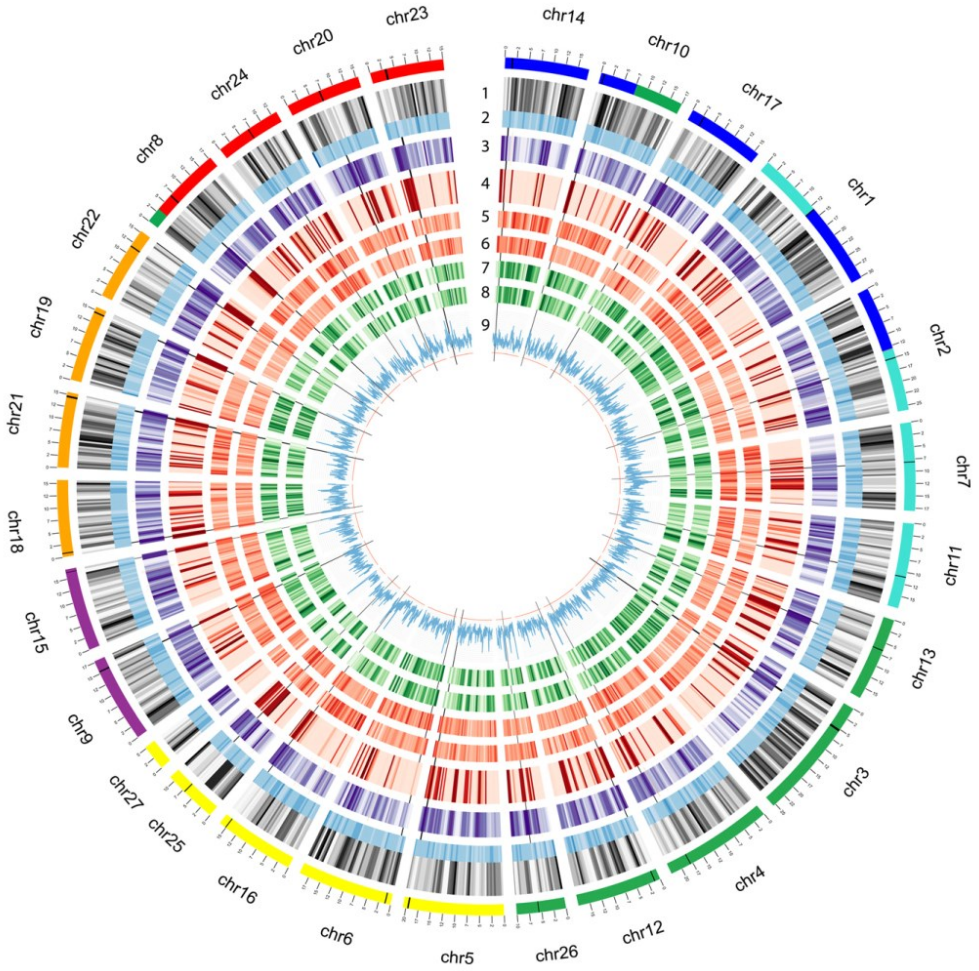



**Figure 4.** Chromosome structure, focus on epigenetic marks.
From outer to inner: karyotype bands colored according to ancestral genome blocks as in Figure 5, followed by: (1) gene density (grey) normalized 0,1; (2) GC content 0.25–0.45 (blue); (3) all TEs density (violet) normalized 0,1, NCLDV evidence is shown as radial orange lines; (4) methylation (red): CHH+CHG+CG, each median per window normalized 0,1, 0.0–3.0 (individual tracks see Figure S32); (5) gametophore H3 repression marks (red, K27me3, K9me2) percent per window normalized, 0.0–2.0 (for more detailed plots see File S1); (6) protonema H3 repression marks (red, K27me3, K9me2) normalized as in (5); (7) gametophore H3 activation marks (green, K4me3, K27Ac, K9Ac) normalized as in (5); (8) protonema H3 activation marks (green, K4me3, K27Ac, K9Ac) normalized as in (5); (9) Nucleotide diversity (blue histogram) 0.0–0.01. Dominant RLC5 peak radius as in Figure 2. (9) 100 kbp sliding window and 100 kbp jump, all other plots as in Figure 2. Chromosomes are arranged according to the ancestral (pre-WGD) seven chromosome karytope (Figure 5).

Such genes are typically constitutively expressed and evolutionarily conserved; however, the functional relevance of GBM in flowering plants remains unclear (Zilberman, 2017). The low incidence of genic methylation in *P. patens*, although all DNA methyltransferase classes are present (Dangwal *et al.*, 2014), probably reflects secondary reduction. Despite the generally low genic methylation, 2012 (5.7%) protein-coding genes contain at least one methylated position in gametophores (Figure S29), and 1155 (3.3%) of the genes show more than 50% of methylatable positions to be methylated (Figure S30), making them GBM candidates. Most methylated genes are not

expressed in gametophores (1608 genes, 79.9%), suggesting that, contrary to flowering plants, GBM might silence them. They are also significantly less often annotated (21.7% of methylated genes carry GO terms, versus 48.7% of all genes; $P < 0.01$, chi-squared test). CHH-type methylation is most abundant (1409 genes), followed by CHG (1306) and CG (1162); one-third of the genes share methylation in all three contexts. The presence of CG methylation in *P. patens* gene bodies is in contrast with a previous report (Bewick *et al.*, 2017), potentially due to different coverage or filtering applied. Surprisingly, given that cytosines are methylated, the average GC content of GBM genes (36.5%) is significantly ($P < 0.01$, T-test) lower than the genome-wide GC (45.9%). Genes without expression evidence in gametohores have lower GC content and GBM than those that are weakly expressed (Table S18, RPKM 0–2), while confidently expressed genes (RPKM >2) are more GC-rich and less methylated. In summary, in contrast with flowering plants low GC genes with no conserved function are principally more often found to be targeted (silenced) by DNA methylation, suggesting their potential conditional activation. GO bias analysis of the methylated genes expressed in gametophores shows enrichment of genes involved in protein phosphorylation (Figure S30(b)). Most (290, 59%) of the expressed methylated genes are expressed in protonema, gametophores and green sporophytes (Figure S30(c)), but 12.5% are expressed in two tissues each, while 17 (3.5%) are exlusively expressed in protonemata, 28 (5.7%) in gametophores and 93 (19%) in green sporophytes.

*Do giant virus remnants guard gametes?* We mapped the genomic segments that were likely acquired horizontally from nucleocytoplasmic large DNA virus relatives [NCLDV, (Maumus *et al.*, 2014); Table S16, and Figures 4 and S14–S22] and found that 87 integrations (NCLDVI) harbor 257 regions homologous to NCLDV protein-coding genes and 163 sRNA clusters. Colinearity and molecular dating analyses of NCLDVIs (Figures S19 and S20) suggest four groups of regions that have been either amplified by recombination events or represent simultaneous integrations. The timing of these integrations (comprising both relatively young and older insertions/duplications) appears independent from the periods of LTR-RT activity. NCLDVI regions are the most variable annotated loci in terms of nucleotide diversity (Figure S18). Previous evidence suggested that NCLDVI represent non-functional, decaying remnants of ancestral infections that are transcriptionally inactivated by methylation (Maumus *et al.*, 2014). By screening available sRNA-seq libraries we could record repetitive, but specific sRNA clusters for these loci. Strikingly, we identified two NCLDV genes harboring sRNA loci that exhibit high transcriptional activity, coinciding with lower levels of DNA methylation as compared with other NCLDVI (Figures S14 and S15). Consistent with the predicted potential to form hairpin structures, sRNA northern blots (Figure S22) of wild type and Dicer-like (DCL) deletion mutants (Khraiwesh *et al.*, 2010; Arif *et al.*, 2012) suggest that RNA transcribed from these loci might be processed by distinct DCL proteins to generate siRNAs. These siRNAs in turn might act to target viral mRNA during a potential NCLDV infection, or to guide DNA methylation to silence these regions (Kawashima and Berger, 2014). Regions harboring corresponding antisense sRNA loci are enriched for stop-codon-free (i.e. non-degrading) NCLDV genes and deviate from the remainder of NCLDVI in terms of cytosine versus histone modifications (Figures S15 and S16). Based on the similarity with intact LTR-RTs in terms of methylation and low GC (Figure S17), and the absence of H3K9me2, we hypothesize that (like intact TEs) these ancient, retained NCLDVi are euchromatic. We propose that they are demethylated during gametogenesis by DEMETER (which in Arabidopsis preferentially targets small, AT-rich, and nucleosome-depleted euchromatic TEs (Ibarra *et al.*, 2012)). Given the proposed time point of activation of these regions during gametangiogenesis, NCLDVIs might provide a means to provide large numbers of siRNAs which, besides ensuring the transgenerational persistence of silencing, could also provide protection against cytoplasmatically replicating viruses *via* RNAi and methylation of the viral genome. This would provide efficient protection for moss gametes which, due to their dependency on water, might be the most exposed to NCLDV infections. This hypothesis provides a plausible answer to the question why endogenous NCLDV relatives have only been found in embryophytes with motile sperm cells (Maumus *et al.*, 2014).

*Genetic variability.* Sequencing three different accessions we find 264 782 SNPs (1 per 1783 bp) for Reute (collected close to Freiburg, Germany), 2 497 294 (1 per 188 bp) for Villersexel (Haute-Saône, France) and 732 288 (1 per 644p) for Kaskaskia (IL, USA) as compared with Gransden. There are 42 490 polymorphisms shared among all three accessions relative to Gransden, with other SNPs present in only one or two of the accessions (Figure S31). SNP densities of *Arabidopsis thaliana* ecotypes occur at one SNP per 149–285 bp (Cao *et al.*, 2011), similar to that in Villersexel, which is surprising given that the rate of neutral mutation fixation is lower in *P. patens* (Rensing *et al.*, 2007). However, Villersexel has an extraordinarily high divergence compared with other *P. patens* accessions (McDaniel *et al.*, 2010). Due to the fact that all accessions are inter-fertile, yet genetically divergent (Beike *et al.*, 2014), and exhibit phenotypic differences (File S2; Hiss *et al.*, 2017), we consider them potential ecotypes. For all accessions, most SNPs (>80%) are found in intergenic and adjacent (potential regulatory) regions of genes (Table S19). Less than 5%

of all SNPs are found in genic regions, of those 34–36% are silent (synonymous), 62–64% missense (non-synonymous) and 1.6% cause a nonsense mutation. Overall, Reute showed 72 regions of SNP accumulation, whereas Villersexel and Kaskaskia showed 30 and 32, respectively (Table S20-S22). The SNP accumulation regions in Reute are more gene-rich with 18 genes/region compared with 8 and 10 in Villersexel and Kaskaskia. One peak on chromosome 16 is found in all accessions and contains genes involved in sterol catabolism and chloroplast light sensing/movement (Figure S33). Sterols have been implicated in cell proliferation, in regulating membrane fluidity and permeability, and in modulating the activity of membrane-bound enzymes (Hartmann, 1998). The over-represented terms detected in the genes commonly harboring SNPs might be the signature of evolutionary modification of dehydration tolerance, for which membrane stability has been shown to be an important factor in mosses (Oliver *et al.*, 2004; Hu *et al.*, 2016).

*Recombination might be needed for purging TEs.* Many genomes have higher densities of TEs in centromeres, sub-telomeres (Figure S34), and sex chromosomes, i.e. regions of low recombination (Dolgin and Charlesworth, 2008). One potential explanation for this biased distribution is that TEs insert with more or less equal frequencies across the genome, but are heterogeneously distributed because purifying selection is weaker in regions of low recombination. This hypothesis can be put to test using the *Physcomitrella* genome: the species is mostly selfing (it practises *de facto* asexual reproduction using sexual gametes; Perroud *et al.*, 2011), and thus the effective rate of recombination is low (since genetic variants are seldom mixed as heterozygotes), and purifying selection is correspondingly weak (Szovenyi *et al.*, 2013). If recombination (in outcrossed offspring) is indeed critical for making purifying selection effective at purging weakly deleterious TEs, we would predict that selection against TE disruption of gene expression may be playing an important role in the chromosomal distribution of TEs (Wright *et al.*, 2003). Hence, the unusual chromosomal structure might be a function of predominant inbreeding. We expect that the genomes of bryophytes that are outcrossers, like *Marchantia polymorpha*, *Ceratodon purpureus*, *Funaria hygrometrica* or *Sphagnum magellanicum*, might show a more biased distribution of TEs along their chromosomes.

**Genome evolution**

*Two whole genome duplication events.* Based on synonymous substitution rates (Ks) of paralogs, at least one WGD event was evident in *P. patens* (Rensing *et al.*, 2007, 2008). However, gene family trees often show nested paralog pairs, and the ancestral moss karyotype is hypothesized to be seven (Rensing *et al.*, 2012), while the extant chromosome number of *P. patens* is $n = 27$ (Reski *et al.*, 1994), suggesting two ancestral WGD events (Rensing *et al.*, 2007, 2012). Using the novel pseudochromosome structure, Ks-based analyses support two WGDs dating back to 27–35 and 40–48 Ma (Figure 5), respectively (cf. supplementary material IV.). Given the detected synteny, the most parsimonious explanation for the extant chromosome number is the duplication of seven ancestral chromosomes in WGD1, followed by one chromosomal loss and one fusion event during the subsequent haploidization. In WGD2 the 12 chromosomes would have duplicated again, followed by five breaks and two fusions, leading to 27 modern chromosomes. The Ks values of the above-mentioned structure-based peaks (Figure 5) fall approximately between 0.5–0.65 (younger WGD2) and 0.75–0.9 (older WGD1). The structural and Ks information can be used to trace those genes that were present in the ancestral (pre-WGD) karyotype and have since been retained (Figure S37 and File S3). In total, 484 genes can be traced to the pre-WGD1 karyotype (denoted ancestor 7), and 3112 genes to the pre-WGD2 karyotype (ancestor 12). GO bias analysis of the ancestor 7 genes shows over-representation of many genes involved in regulation of transcription and metabolism (Figure S38). This accords with previous evidence that metabolic genes were preferentially retained after the *P. patens* WGD (Rensing *et al.*, 2007), and with the trend that genes involved in transcriptional regulation are preferentially retained after plant WGDs (De Bodt *et al.*, 2005).

*WGDs are common in mosses, but not in other bryophytes.* Detecting WGD events using paranome-based Ks distributions is notoriously difficult (Vekemans *et al.*, 2012; Vanneste *et al.*, 2014). Here we compared several methods for deconvolution of such distributions and found that a mixture model based on log-transformed values was able to detect four potential WGDs (Figure S39), including the two that we observed based on the pseudochromosomal structure (Figure 5). By excluding very young/low and very old/high Ks ranges, we restricted the data to the two structure-based events. Using low bandwidth (smoothing) we find that such methodology is able to detect relatively young WGDs with a clear signature (Figure S39(e, f)), whereas overlapping distributions (here the older WGD1) are hinted at via significant changes in the distribution curve at higher bandwidth settings (Figure S39(i, j); cf. Experimental Procedures and Appendix S1 Supplementary Material IV/2 for details). We applied this paranome-based WGD prediction to transcriptome data obtained from the onekp project (www.onekp.com) on 41 moss, 7 hornwort and 28 liverwort datasets and overlaid them with a molecular clock tree (Figures S40–S42) (Newton *et al.*, 2006). For 24 of the moss samples at least one WGD signature was supported. For four out of these 24

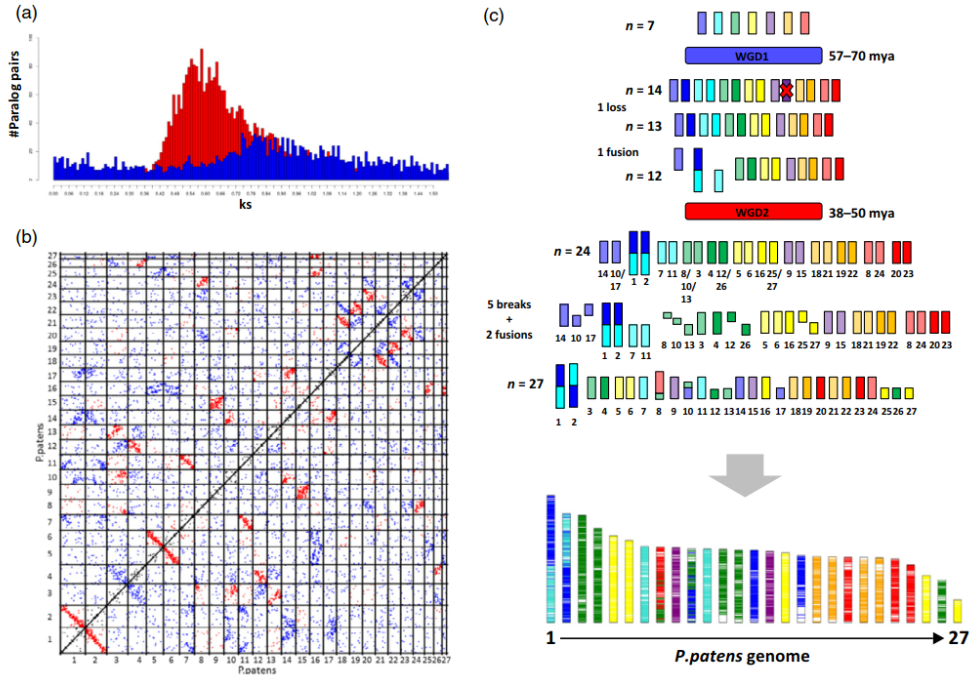



**Figure 5.** Evolutionary scenario leading to the modern *P. patens* genome.
(a) Ks distribution (*y*-axis) of paralogous pairs (*x*-axis) inherited from two (blue for older and red for more recent) WGD events.
(b) Dotplot representation of the paralogous pairs belonging to two WGD events.
(c) Karyotype evolution of the *P. patens* genome from an $n = 7$ ancestor through two WGDs. The modern *P. patens* genome is illustrated as a mosaic of coloured chromosomal blocks highlighting chromosome ancestry.

moss datasets, mixture model components were merged into one WGD signature with the possibility of additional hidden WGD signatures. Among these species is *Physcomitrium* sp. which is a close relative of *P. patens*; shared WGD events are in accordance with previous studies (Beike *et al.*, 2014). The three *Sphagnum* species show overlap and significant gradient change support for a young WGD event and in *Sphagnum lescurii* also significant support for an older WGD event, supporting a recent report (Devos *et al.*, 2016). While only a chromosome-scale assembly would be able to detect WGD events with high confidence, we note that evidence of WGDs is not detected in any of the liverwort and hornwort datasets, while the majority of moss lineages appears to have been subject to ancient WGDs. In contrast with mosses (Rensing *et al.*, 2012; Szovenyi *et al.*, 2014), most liverworts and are known for low levels of neopolyploidy and endopolyploidy with rather constant chromosome numbers within each lineage (Bainard *et al.*, 2013). The three-fold fluctuations in genome size in nested hornwort lineages without a chromosomal change (Bainard and Villarreal, 2013) is thus most likely due to variable TE content. The karyotype evolution of *P. patens* can thus be considered as typical for moss genomes, but probably different from the genomes of hornworts and liverworts. While we do not know why mosses might be more prone to fixation of genome duplications than other bryophytes, the associated paralog acquisition and retention might be a foundation for the relative species richness of mosses (Rensing, 2014; Rensing *et al.*, 2016; Van de Peer *et al.*, 2017).

*Ancient colinearity reveals conserved plant-specific functions.* Have gene orders been conserved since the last common ancestor of land plants (LAP)? Colinearity analyses with 30 other plant genomes (cf. Experimental Procedures and Appendix S1 Supplementary Material IV/3) revealed 180 colinear regions, harbouring around 1700 genes. *P. patens* chromosomes contain 0.5–10 of these genes per Mbp (Figure S43), most chromosomes hence containing a number of syntenic genes that follows

random expectation. Chromosomes 1, 8, 11, 14, 16 and 27, however, contain significantly more ancient colinear genes than expected ($q < 0.05$, Fisher's exact test; File S3). GO bias analyses revealed that chromosome 8 is enriched for genes encoding functions for plant cell and tissue growth and development (Figure S44). Surprisingly, several hundred genes are present in colinear regions that involve 5–21 other species. Moreover, 17 of these regions showed elevated levels of gene co-expression ($P < 0.05$, permutation statistics; File S3), indicating potential co-regulation of neighboring genes, thus corroborating the existence of conserved plant regulons (Van de Velde *et al.*, 2016) or genomic regions exposed similarly to the transcriptional machinery. GO bias analyses of these ancient syntenic genes demonstrate that they are involved in land plant-specific cell growth and tissue organization (Figure S45), akin to chromosome 8. Apparently, genes encoded in the LAP genome that enabled the distinct cell and tissue organization of land plants have been retained as colinear blocks throughout land plant evolution. In total, 10 genes on chromosome 7 can be traced back to chromosome 4 of ancestor 12 (pre-WGD2), and to chromosome 2 of ancestor 7 (pre-WGD1). GO bias of chromosome 7 (Figure S46) further supports the notion that genes enabling plant-specific development have been conserved since the LAP.

## CONCLUSIONS

Our analyses show that the genome of the model moss is organized differently from seed plant genomes. In particular, no central TE-rich and distal gene-rich chromosomal areas are detected, and centromeres are potentially marked by a subclass of Copia elements. There is evidence for activation of TE and viral elements during the life cycle of *P. patens* that might be related to its haploid-dominant life style and motile gametes. Surprisingly, syntenic blocks harboring genes involved in plant-specific cell organization were conserved for *ca.* 500 Ma of land plant evolution. Chromosome-scale assemblies of other non-seed plants will be needed in order to understand how plant genomes from diverse lineages evolve, and to determine whether the genomes of haploid-dominant plants are generally different from those of seed plants.

## EXPERIMENTAL PROCEDURES

### Sequencing and assembly

We sequenced *Physcomitrella patens* Gransden 2004 using a whole genome shotgun sequencing strategy. Most sequencing reads were collected with standard Sanger sequencing protocols on ABI 3730XL capillary sequencing machines at the Department of Energy Joint Genome Institute in Walnut Creek, California, USA (http://www.jgi.doe.gov/sequencing/protocols/prots_production.html) as previously reported (Rensing *et al.*, 2008). BAC end sequences were collected using standard protocols at the HudsonAlpha Institute in Huntsville, Alabama, USA. The sequencing (see Table S1) consisted of two libraries of 3 kbp (4.01x), 3 libraries of 8 kbp (4.58x), four fosmid libraries (0.43x), and two BAC libraries (0.22x) on the Sanger platform for a total of 9.25x Sanger based coverage. In total, 7 572 652 sequence reads (9.25x assembled sequence coverage, see Table S1 for library size summary) were assembled using our modified version of Arachne v.20071016 (Jaffe *et al.*, 2003) with parameters correct1_passes=0 maxcliq1 = 140 BINGE_AND_PURGE=True max_bad_look=2000 (see Table S2 for overall scaffold and contigs statistics). This produced a raw assembly consisting of 1469 scaffolds (4485 contigs) totaling 475.8 Mb of sequence, with a scaffold N50 of 2.8 Mb, 271 scaffolds larger than 100 kbp (464.3 Mb). Scaffolds were screened against bacterial proteins, organellar sequences and the GenBank 'nr' database, and removed if found to be a contaminant. Additional scaffolds were removed if they were: (i) scaffolds smaller than 50 kbp consisting of >95% 24-mers that occurred four other times in scaffolds larger than 50 kbp; (ii) contained only unanchored RNA sequences; (iii) were less than 1 kbp in length; or (iv) contaminated. Post-screening, we integrated the resulting sequence with the genetic map reported here (3712 markers), and BAC/fosmid paired end link support. An additional map (9080 markers) was developed for chromosome 16 that resolved ordering problems present in the original map, and was used for the integration of chromosome 16. The integrated assembly was screened for contamination to produce a pseudomolecule reference covering 27 nuclear chromosomes. The pseudomolecules include 462.3 Mb of base pairs, an additional 351 unplaced scaffolds consist of 4.9 Mb of unanchored sequence. The total release includes 467.1 Mb of sequence assembled into 3077 contigs with a contig N50 of 464.9 kbp and an N content of 1.5%. Chromosome numbers were assigned according to the physical length of each linkage group (1 = largest and 27 = smallest).

### Genetic mapping

In order to assign the sequenced scaffolds representing the release version V1.2 *Physcomitrella* genome sequence to chromosomes, we used a genetic mapping approach based on high-density SNP markers. SNP loci between the Gransden 2004 ('Gd') and genetically divergent Villersexel K3 ('Vx') genotype were identified by Illumina sequencing (100 bp end reads; Illumina GAII) of the Vx accession. The sequence data have been deposited in the NCBI Sequence Read Archive as accessions SRX037761 (two Illumina Genome Analyzer II runs: 176.1 M spots, 26.8 G bases, 93.4 Gb downloads) and SRX030894 (three Illumina Genome Analyzer II runs: 277.9 M spots, 42.2 G bases, 56 Gb downloads). SNPs for linkage mapping were selected for the construction of an Illumina Infinium bead array for the GoldenGate genotyping platform, based on their distribution across the 1921 scaffolds representing the V1.2 genome sequence assembly, with an average physical distance between SNP loci of *ca.* 110 kbp. Segregants of a mapping population [539 progeny from Gd×Vx crosses: (Kamisugi *et al.*, 2008)] were genotyped at 5542 loci to construct a linkage map using JoinMap 4.0 (Van Ooijen JW, 2006, Kyazma B.V., Wageningen, The Netherlands), with a minimum independence LOD threshold of 22, a recombination threshold of 0.4, a ripple value of 1, a jump threshold of 5 and Haldane's mapping function. Of the 5542 SNPs, 4220 loci were represented in the final map. The map contained 27 linkage groups, covering 5432.9 cM. Map lengths were calculated using two methods: one in which L (total map length) = Σ [(linkage group length) + 2 (linkage group length/no. markers)] (Fishman *et al.*, 2001) and one in which L = Σ[(linkage group length (no. markers + 1)/(no. markers − 1)] (Chakravarti *et al.*, 1991). The map corresponded to 467 985 895 bp distributed

across the previously predicted 27 *P. patens* chromosome (Table S3). Chromosome numbers were assigned according to the overall physical length of each linkage group (1 = largest and 27 = smallest).

### Pseudochromosome construction

The combination of the existing genetic map (4220 markers), and BAC/fosmid paired end link support was used to identify 12 misjoins in the overall assembly. Misjoins were identified as linkage group discontiguity coincident with an area of low BAC/fosmid coverage. In total, 12 breaks were executed, and 295 scaffolds were oriented, ordered and joined using 268 joins to form the final assembly containing 27 pseudomolecule chromosomes, capturing 462.3 Mb (98.97%) of the assembled sequence. Each chromosome join is padded with 10 000 Ns. The final assembly contains 378 scaffolds (3077 contigs) that cover 467.1 Mb of the genome with a contig L50 of 464.9 kbp and a scaffold L50 of 17.4 Mb.

Completeness of the euchromatic portion of the genome assembly was assessed using 35 940 full-length cDNAs. The aim of this analysis was to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. The cDNAs were aligned to the assembly using BLAT (Kent, 2002); Parameters: –t=dna –q=rna –extendThroughN, and alignments ≥90% bp identity and ≥85% coverage were retained. The screened alignments indicate that 34 984 (97.3%) of the FLcDNAs aligned to the assembly. The ESTs that failed to align were checked against the NCBI nucleotide repository (nr), and a large fraction was found to be prokaryotic in origin. Significant telomeric sequence was identified using the TTTAGGG repeat, and care was taken to make sure that it was properly oriented in the production assembly. Plots of the marker placements for the 27 chromosomes are shown in File S2. For contamination screening, further assessment of assembly accuracy and organellar genomes please refer to Appendix 1, Supplementary Material, Section I.

### Mapping of the v1.6 genome annotation

Gene models of the v1.6 annotation (Zimmer *et al.*, 2013) were mapped against the V3 assembly using GenomeThreader (Gremme *et al.*, 2005) and resulting spliced alignments were filtered and classified for consistency with the original gene structures. 93.9% of the 38 357 v1.6 transcripts could be mapped with unaltered gene structure. This comprised 29 371 loci (91.4% of the v1.6 loci). The majority of the unmappable v1.6 models represented previously unidentified bacterial or human contaminations in the V1 assembly (492 loci). Nevertheless, 49 loci with expression evidences remained unmappable in the current assembly. The mapped annotation is made available via the cosmoss.org genome browser and under the download section.

### Generation of the v3.1 genome annotation

All available RNA-seq libraries (File S3 and Table S10) were mapped to the V3 assembly using TopHat (Trapnell *et al.*, 2009). Based on a manually curated set of cosmoss.org reference genes (Zimmer *et al.*, 2013), libraries and resulting splice junctions were filtered to enrich evidence from mature mRNAs. Sanger and 454 EST evidence used in the generation of the v1.6 annotation was mapped using GenomeThreader. The resulting splice junctions and exonic features were used as extrinsinc evidences to train several gene finders, which were evaluated using the cosmoss.org reference gene set. Based on this evaluation, five predictive models derived with EuGene (Foissac *et al.*, 2003) resulting from different parameter combinations, including the original model used to predict v1.6, were retained for genome-wide predictions. RNA-seq libraries were assembled into virtual transcripts using Trinity (Grabherr *et al.*, 2011). The resulting 1 702 106 assembled transcripts with a mean length of 1219 bp were polyA trimmed using seqclean (part of the PASA software), of which 96% could be mapped against the V3 genome using GenomeThreader. Together with the 454 and Sanger ESTs 2 755 148 transcript sequences were used as partial cDNA evidence in the PASA software to derive 266 051 assemblies falling in 68 382 subclusters. For these, transdecoder was trained and employed to call open reading frames based on PFAM (Finn *et al.*, 2016) domain evidence. Gene models from transdecoder, EuGene and the JGI V3.0 predictions were combined and evaluated using the eval software (Keibler and Brent, 2003) on the reference gene set. Based on the resulting gene and exon sensitivity and specificity scores a rank-based weight was inferred (Table S9), which was used to infer combined CDS models using EVidenceModeler, resulting in a gene sensitivity/specificity of 0.76/0.76 and an exon sensitivity/specificity of 0.93/0.98. For these combined CDS features, UTR regions were annotated using PASA in six iterations. All transcript evidence and alternative gene models are available via tracks in the cosmoss.org genome browser. From the resulting set of gene models, protein-coding gene loci and representative isoforms were inferred using a custom R script implementing a multiple feature weighting scheme that employed information about CDS orientation, proteomic, sequence similarity and expression evidence support, feature overlaps, contained repeats, UTR-introns and UTR lengths of the gene models in a Machine Learning-guided approach. This approach was optimized and trained based on a manually curated training set in order to ideally select the functional, evolutionary conserved 'major' isoform for each protein-coding gene locus. The v3.1 annotation comprises only the 'major' (indicated by the isoform index 1 in the CGI), while v3.3 also includes other splice variants with isoform indices >1.

### Availability of gene models and additional data

The analyses in this publication rely on the structural annotation v3.1. Subsequently, this release was merged with the phytozome-generated release v3.2, leading to the current release v3.3 which is available from http://cosmoss.org and https://phytozome.jgi.doe.gov/. Both v3.1 and v3.3 are available in CoGe (https://genomevolution.org/coge/GenomeView.pl?gid=33928), and v1.6 and v1.2 can be loaded as tracks for backward compatibility. Available experiment tracks can be downloaded and are listed in Table S12. Organellar genomes are also available at CoGe under the id 35274 (chloroplast) and 35275 (mitochondrion). For gene annotation version 3.2/3.3, locus naming, non-protein coding genes and functional annotation refer to Appendix S1, Supplementary Material, Section II. Annotations v3.1 and v3.3 are available in File S1, including a lookup of gene names for versions 3.3, 3.1, 1.6, 1.2 and 1.1. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession ABEU00000000. The version described in this paper is version ABEU02000000.

*Cytological analyses.* The chromosome arrangement during mitotic metaphase as well as the punctate labelling at pericentromeric regions after immunolabelling with a pericentromere-specific antibody against H3S28ph (Gernand *et al.*, 2003) indicate a monocentric chromosome structure in *P. patens* (Figure S5). Furthermore, many plant genomes, as for example *A. thaliana* (Fuchs *et al.*, 2006), are organized in well defined heterochromatic pericentromeric regions, decorated with typical heterochromatic marks

(H3K9me1, H3K27me1) and gene-rich regions presenting the typical euchromatic marks (H3K4me2). By contrast, immunostaining experiments with antibodies against these marks label the entire chromatin of flow-sorted interphase *P. patens* nuclei homogeneously (Figure 3(b)). Obviously, *P. patens* nuclei are thus characterized by a uniform distribution of euchromatin and heterochromatin.

### Transposon and repeat detection and annotation

TRharvest (Ellinghaus *et al.*, 2008) which scans the genome for LTR-RT specific structural hallmarks (like long terminal repeats, tRNA cognate primer binding sites and target site duplications) was used to identify full length LTR-RTs. The input sequences comprised the 27 pseudochromosomes plus all genomic scaffolds with a length of ≥10 kbp together with a non-redundant set of 183 *P. patens* tRNAs, identified beforehand via tRNA scan (Lowe and Eddy, 1997). The used parameter settings of LTRharvest were: 'overlaps best -seed 30 -minlenltr 100 -maxlenltr 2000 -mindistltr 3000 -maxdistltr 25000 -similar 85 -mintsd 4 -maxtsd 20 -motif tgca -motifmis 1 -vic 60 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3'. All of the resulting 9290 candidate sequences were annotated for PfamA domains with hmmer3 (http://hmmer.org/) and stringently filtered for false positives by several criteria, the main ones being the presence of at least one typical retrotransposon domain (e.g. RT, RH, INT, GAG) and a tandem repeat content below 25%. The filtering steps led to a final set of 2785 high confident full-length LTR RTs. Transposons were annotated by RepeatMasker (Smit *et al.*, 1996) against a custom-built repeat library (Spannagl *et al.*, 2016) which included *P. patens* specific full length LTR-retrotransposons.

Repetitive elements have also been annotated *de novo* with the REPET package (v2.2). The TEdenovo pipeline from REPET (Flutre *et al.*, 2011) was launched on the contigs of size >350 kbp in the v3 assembly (representing approximately 310 Mb, gaps excluded) to build a library of consensus sequences representative of repetitive elements. Consensus sequences were built if at least five similar hits were detected in the sub-genome. Each consensus was classified with PASTEC (Hoede *et al.*, 2014) followed by semi-manual curation. The library was used for a first genome annotation with the TEannot pipeline (Quesneville *et al.*, 2005) from REPET to select the consensus sequences that are present for at least one full length copy ($n$ = 349). Each selected consensus was then used to perform final genome annotation with TEannot with default settings (BLASTER sensitivity set to 2). The REPET annotations absent from the mipsREdat annotation were added to the latter to build the final repeat annotation. Tandem repeats Finder (Benson, 1999) was launched with the following suite of parameters: 2 7 7 80 10 50 2000. The putative centromeric repeat previously identified through tandem repeats analysis (Melters *et al.*, 2013) was compared with the whole V3 assembly using RepeatMasker (Smit *et al.*, 1996) with default settings (filter divergence <20%). Besides Copy and Gypsy-type elements (see main text), other types of TEs, including LINEs and Class II (DNA transposon) elements, appear at very low frequency (0.1% each). Simple sequence repeats represent only 2% of the assembly. For TE phylogenetic, age and expression analyses as well as NCLDV analyses refer to Appendix S1, Supplementary Material, Section III.

### ChIP-seq data

Published CHIP-seq data (Widiez *et al.*, 2014) for *P. patens* were re-analysed by mapping read libraries against the *P. patens* V3.0 genome sequence. Briefly, the FASTA and QUAL files were converted into FASTQ data files, which were aligned against the *P. patens* v3.0 genome using BWA v0.5.9 (Li and Durbin, 2010), employing a seed length of 25, allowing a maximum of two mismatches on the seed and a total maximum of 10 mismatches between the reference and the reads. In order to avoid redundancy problems, all reads that were mapped to more than one genomic locus were omitted as already applied elsewhere (Zemach *et al.*, 2010; Stroud *et al.*, 2012). SAM files were converted into BED files using an in-house Python script.

### Identification of histone-modified enriched regions

For the identification of the histone-modified enriched regions (peaks) the software MACS2 v2.0.10 (Zhang *et al.*, 2008; Feng *et al.*, 2012) with parameters tuned for histone modification data was used. The parameters used were 'no model', shift size set as 'sonication fragment size', 'no lambda', 'broad', bandwidth 300 following the developer's instructions, fold change between 5 and 50 and q-value 0.01. As control for the peak identification the combination of Input-DNA and Mock-IP of the corresponding tissues was used as in Widiez *et al.* (2014). The number of identified peaks per tissue and histone mark is shown in Table S17.

### Extension of unannotated genomic regions

For several gene models in the *P. patens* v3.1 genome annotation the prediction of UTR regions (either 5′ or 3′) failed. In total there are 9769 genes lacking the 5′-UTR and 11 385 genes lacking the 3′-UTR. Additionally, gene promoters are also unannotated. Using an approach already used in (Widiez *et al.*, 2014), UTRs and promoters were assigned to gene models. In brief, a Python script was implemented that takes as input any valid GFF3 file and: (i) creates UTR regions of 300 bp for genes lacking either one or both of them; and (ii) creates potential promoter regions of 1500 bp upstream and downstream of each gene in the file. In the case that the space between the gene and the next element is not wide enough for the extension of the gene model by 300 bp, the new UTR region is shrunk to the available space. In the case that two consecutive genes have to be extended and the space between them is less than 2 × 300 bp the new UTRs are assigned half the space between the two genes. For the assignment of promoters the same rules apply. In no case is an element created that overlaps with existing elements of the annotation file used as input.

### Filtering for expressed genes

Based on all the available JGI gene atlas (http://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/) RNA-seq data downloaded from Phytozome (File S3), we filtered for genes that had a certain minimal RPKM value in at least one condition. At RPKM 2, 20 274 genes are expressed, at RPKM 4 18 281 genes. The RPKM cutoff of four was based on quantitative real-time PCR (qRT-PCR) results of a recent microarray transcriptome atlas study (Ortiz-Ramirez *et al.*, 2015), in which genes with this expression level were reliably detected by qPCR.

### BS-seq data: plant material and culture conditions

*Physcomitrella patens* accession Gransden was grown in 9-cm Petri dishes on 0.9% agar solidified minimal (Knop's) medium. Cultures were grown under the following experimental conditions: 16 h/8 h light/dark cycle, 70 μmol $sec^{-1}$ $m^{-2}$, for 6 weeks at 22°C/19°C day/night temperature following 8 h/16 h light/dark cycle, 20 μmol $sec^{-1}$ $m^{-2}$, for 7 weeks at 16°C/16°C day/night temperature. Adult gametophores were harvested after 13 weeks and DNA was isolated according to Dellaporta *et al.* (1983) with minor modifications (Hiss *et al.*, 2017).

### Bisulfite conversion, library preparation and sequencing

Bisulfite conversion and library preparation was conducted by BGI-Shenzen, Shenzen, China according to the following procedure: DNA was fragmented to 100–300 bp by sonication, followed by blunt end DNA repair adding 3′-end dA overhang and adapter ligation. The ZYMO EZ DNA Methylation-Gold kit was used for bisulfite conversion and after desalting and size selection a PCR amplification step was conducted. After an additional size selection step the qualified library was sequenced using an Illumina GAII instrument according to manufacturer instructions resulting in 66 108 645 paired end reads of 90 bp length.

### Processing of BS-seq reads

Trimmomatic v0.32 (Bolger *et al.*, 2014) was used to clean adapter sequences, to trim and to quality-filter the reads using the following options: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:4:5 TRAILING:3 MINLEN:35 resulting in cleaned paired-end and orphan single-end reads. Further, the paired-end and single-end reads were mapped with Bismark v0.14 (Krueger and Andrews, 2011) against *P. patens* chloroplast (NC_005087.1) and mitochondrion (NC_007945.1) sequences using the *–non_directional* option due to the nature of the library. After mapping the remaining single-end and paired-end reads with Bismark v0.14 separately against the genome of *P. patens* both SAM alignment files were sorted and merged with samtools v0.1.19 (Li *et al.*, 2009) and deduplicated with the *deduplicate_bismark* program of Bismark v0.14. To call methylation levels for the different cytosine contexts (CG, CHG, CHH), deduplicated SAM files and the R package *methylkit* (Akalin *et al.*, 2012) were used, only considering sites with a coverage of at least nine reads and a minimal mapping quality of 20.

### Gene- and TE-body methylation

Gene- and TE-body methylation levels were calculated for individual cytosine contexts (CG, CHG, CHH). For each gene and TE, all annotated feature regions (promoter, 5′-UTR, CDS, intron, 3′-UTR, TE-fragment) were combined and divided into 10 quartiles. For each quartile the mean methylation level (CG, CHG, CHH) was calculated and the average, 5% and 95% distribution per quartile and feature type were plotted. For the TE-body methylation plots TEs were further subdivided into TE-groups. For gene body methylation (GBM) analysis positions were filtered according to ≥90% of the reads showing methylation. Distribution of affected genes over the three different contexts was analysed with Venny (Figure S29; http://bioinfogp.cnb.csic.es/tools/venny/) and visualized via a stacked column diagram (Figure S30). Genes were grouped by RPKM value (0;>0 < 2;≥2) and compared with regard to GC and methylation content (Table S18).

### Read mapping and variant calling

Genomic DNA sequencing data for *P. patens* accessions Reute (SRP068341), Villersexel (SRX030894) and Kaskaskia (SRP091316) are available from the NCBI Sequence Read Archive (SRA). The libraries were trimmed for adapters and quality filtered using trimmomatic v32 (Bolger *et al.*, 2014) applying the following parameters: -phred33 ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:8:5 SLIDINGWINDOW:4:15 TRAILING:15 MINLEN:35. After trimming, the single-end and paired-end reads were initially mapped to the chloroplast genome (NC_005087.1), the mitochondrial genome (NC_007945.1) and ribosomal DNAs (HM751653.1, X80986.1, X98013.1) using GSNAP v2014-10-22 (Wu *et al.*, 2016) with default parameters. The remaining unmapped single-end and paired-end reads were used for reference mapping using GSNAP with default parameters and both resulting SAM alignment files were sorted and merged with samtools v0.1.19 (Li *et al.*, 2009). Duplicated reads were further removed with *rmdup* from samtools to account for potential PCR artifacts. GATK tools v3.3.0 (McKenna *et al.*, 2010) were used for SNP calling as recommended by the Broad institute for species without a reference SNP database including the 'ploidy 1' option for the first and second haplotype calling step.

### SNP validation

Called SNPs of the accession Villersexel were validated by comparing them to the Illumina Infinium bead array dataset (File S3) used for map construction (see Map construction method section). The 4650 bead array probes were mapped to the genome using GSNAP (Wu *et al.*, 2016) and SNPs were called using mpileup and bcftools. In total, 4628 SNPs could be unequivocally mapped, out of those 4466 (96%) were also called as SNPs in the gDNA-seq based Villersexel GSNAP/GATK dataset. Thus, the vast majority of SNPs called based on deep sequence data could be independently confirmed (File S3).

### SNP divergence estimates

To obtain window-wise (100 kbp non-overlapping windows) nucleotide diversity pi and Tajima's *D* values, a 'pseudogenome' was constructed for each accession using a custom python script. In brief, based on the VCF file output generated by GATK all given variants were reduced to SNPs and InDels and for each accession (Kaskaskia, Reute and Villersexel) the corresponding reference sequence was substituted with the ALT allele at the given positions. These 'pseudogenome' FASTA files were additionally masked for all sites which had a read coverage <5 which might lead to erroneous SNP calling. The masked 'pseudogenome' FASTA files were further converted into PHYLIP format and used as input for Variscan v2.0 (Hutter *et al.*, 2006), settings 'RunMode = 12', 'Sliding Window = 1; WidthSW = 100 000; JumpSW = 100 000; WindowType = 0' and excluding alignment gaps via 'CompleteDeletion = 1' (Figure S32).

### SNP accumulation detection

Window-wise (50 kbp with 10 kbp overlap) SNP numbers were extracted from the 'pseudogenome' FASTA files by a custom R script. The R functions fisher.test and p.adjust (method = ' were used to select fragments that show a significantly (adjusted *P*-value <0.01) higher SNP number than the chromosome average. A region of accumulated SNPs (hotspot) was called if at least five adjacent fragments showed a significantly higher SNP number (Tables S20–S22 and Figure S33).

### Structure-based ancestral genome reconstruction and associated karyotype evolutionary model

The *P. patens* genome was self-aligned to identify duplicated gene pairs following the methodology previously described (Salse *et al.*, 2009). Briefly, gene pairs are identified based on blastp alignment using CIP (cumulative identity percentage) and CALP (cumulative alignment length percentage) filtering parameters with respectively 50% and 50%. Ks (rate of synonymous substitutions) distribution of the identified pairs unveiled two peaks illuminating two WGDs, one older and one more recent, included between Ks 0.75–0.9 (WGD1) and 0.5–0.65 (WGD2).

We performed a classical dating procedure of the two WGD events based on the observed sequence divergence, taking into account the Ks ranges between 0.75–0.9 and 0.5–0.65 and a mean

substitution rate (r) of $9.4 \times 10^{-9}$ substitutions per synonymous site per year (Rensing *et al.*, 2007). The time (*T*) since gene insertion is thus estimated using the formula $T = Ks/2r$.

Mapping of the identified gene pairs on the *P. patens* chromosomes defines seven independent (non-overlapping) groups (or CARs for Contiguous Ancestral Regions) of four duplicated regions (representing two rounds of WGDs; Figure S37). Based on the seven CARs identified, we determined the most likely evolutionary scenario based on the assumption that the proposed evolutionary history involves the smallest number of shuffling operations (including inversions, deletions, fusions, fissions, translocations) that could account for the transition from the reconstructed ancestral genome to modern karyotype (Salse, 2012). The ancestor 7 and 12 genes were mapped to the extant chromosomes and visualized as circular plots (Figure S37). These two ancestors (7 and 12) correspond respectively to the pre-WGD1 ancestor (quadruplicated by WGD1 and WGD2 in the modern *P. patens* genome), and the pre-WGD2 ancestor that is the result of the duplication of ancestor 7 (leading to ancestor 14) after one fusion and one chromosome loss (duplicated by WGD2 in the modern *P. patens* genome).

### Paranome-based WGD prediction

For species samples and Ks distribution calculation refer to Appendix 1, Supplementary Material, Section IV. We employed mixture modeling to find WGD signatures using the *mclust* v5.1 R package to fit a mixture model of Gaussian distributions to the raw Ks and log-transformed Ks distributions. All Ks values ≤0.1 were excluded for analysis to avoid the incorporation of allelic and/or splice variants and to prevent the fitting of a component to infinity (Schlueter *et al.*, 2004; Vanneste *et al.*, 2015), while Ks values >5.0 were removed because of Ks saturation. Further, only WGD signatures were evaluated between the Ks range of 0.235 (12.5 Mya) to account for recently duplicated gene pairs to Ks of 2.0 to account for misleading mixture modeling above this upper limit (Vanneste *et al.*, 2014, 2015). Because model selection criteria used to identify the optimal number of components in the mixture model are prone to overfitting (Vekemans *et al.*, 2012; Olsen *et al.*, 2016) we also used SiZer and SiCon (Chaudhuri and Marron, 1999; Barker *et al.*, 2008) as implemented in the *feature* v1.2.13 R package to distinguish components corresponding to WGD features at a bandwidth of 0.0188, 0.047, 0.094 and 0.188 (corresponding 1, 2.5, 5 and 10 Mya) and a significance level of 0.05.

Deconvolution of the overlapping distributions that can be derived from paranome-based Ks values without structural information shows that using mixture model estimation based on log-transformed Ks values mimics structure-based WGD predictions better than using raw Ks values, resulting however in the prediction of four WGD signatures (pbSIG1: 0.15–0.32; pbSIG2: 0.48–0.60; pbSIG3: 0.7–1.12; pbSIG4: 1.66–3.45; Figure S39(a, b)). As WGD signature prediction based on paranome-based Ks values can be misleading and is prone to overprediction (Schlueter *et al.*, 2004; Vekemans *et al.*, 2012; Vanneste *et al.*, 2015; Olsen *et al.*, 2016) we only considered Ks distribution peaks in a range of 0.235–2.0 as possible WGD signatures, thus excluding young paralogs potentially derived from tandem or segmental duplication and those for which accurate dating cannot be achieved due to high age. The paranome-based WGD signatures pbSIG2 (25–32 Ma) overlaps with the younger WGD2, and pbSIG3 (37–60 Ma) overlaps with the older WGD1. Further testing for significant gradient changes in the Ks distribution applying different bandwidths showed that only pbSIG2 is detected as a significant WGD signature (significance level 0.05; Figure S39(h)), whereas pbSIG3 overlaps with a significant change of the Ks distribution curve at a bandwidth of 0.047 but shows no significant gradient change. These results show that even if one paranome-based WGD signature can be found which perfectly overlaps with a structure-based WGD signature (WGD1 and pbSIG3) it is still hard to significantly distinguish it from the younger WGD signatures (WGD2 and pbSIG2) which tend to collapse using higher bandwidths (Figure S39(i, j)). Showing that log-transformed Ks value mixture modeling at least can predict young WGD signatures and can pinpoint older WGD signatures, we applied paranome-based WGD prediction to transcriptome data obtained from the onekp project (www.onekp.com) on 41 moss samples, 7 hornwort samples and 28 liverwort samples and overlaid them with an existing time tree (Figures S40–S42). After evaluating the overlap of significant gradient changes on mixture model components, for 24 out of 41 moss samples at least one WGD signature was supported. For four out of these 24 moss samples mixture model components were merged into one WGD signature with the possibility of additional hidden WGD signatures. Among these samples is *Physcomitrium* sp. which belongs like *P. patens* to the Funariaceae with WGD signatures 3 (0.43–0.66) and 4 (0.80–1.07), overlapping with pbSIG2 and pbSIG3 from *P. patens* and hinting at WGD events in *Physcomitrium* 23–35 Ma and 43–57 Ma ago, respectively. For all liverwort samples and almost all hornwort samples no single predicted WGD signature was supported by three different bandwidth kernel densities. For one hornwort, namely *Megaceros flagellaris*, one WGD signature was supported by a significant gradient change (significance level 0.05), which disappeared using a more stringent significance level of 0.01 and represents more likely a mixture model artifact than a true WGD signature.

### Colinearity analyses

For set of species refer to Appendix S1, Supplementary Material, Section IV. Initially, all chromosomes from all species were compared against each other and significant colinear regions are identified. To detect colinearity within and between species i-ADHoRe 3.0 was used (Proost *et al.*, 2012) with the following settings: alignment_method gg2, gap_size 30, cluster_gap 35, tandem gap 30, q_value 0.85, prob_cutoff 0.01, multiple_hypothesis_correection FDR, anchor_points 5 and level_2_only false. *P. patens* v3.1 genes were assigned to PLAZA 3.0 gene families based on the family information for the best BLASTP match (27 895 genes were assigned to 10 153 gene families). The profile-based search approach of i-ADHoRe combines the gene content information of multiple homologous genomic regions and therefore allows detection of highly degenerated though significant genomic homology (Simillion *et al.*, 2008). In total, 180 regions were found showing significant colinearity with genomes from flowering plants (colinearity with green algal genomes was not found), comprising 1717 genes involved in syntenic regions, representing 660 unique conserved moss genes. Whereas 94/180 of the ultra-conserved colinear (UCC) regions showed genomic homology with one other species, 45 UCC regions showed colinearity with five or more other plant genomes. One UCC region (multiplicon 1440, File S3) grouped 27 genomic segments from 21 species showing colinearity, while 70% of the UCC regions contained five or more conserved moss genes. Starting from the V1 moss genome assembly, only 11/180 UCC regions were recovered, demonstrating that the superior assembly V3 significantly improves the detection of ancient genomic homology. Mapping of the 660 UCC genes reveals their chromosomal location (Figure S43). Co-expression analysis of neighboring UCC genes was performed using the Pearson Correlation Coefficient (PCC) on the JGI gene atlas data (File S3) and permutation

statistics were used to identify UCC regions showing significant levels of gene co-expression (i.e. based on 1000 iterations, in how many cases was the expected median PCC for n randomly selected genes larger than the observed median PCC for n UCC genes).

We tested whether the actual number of genes detected to be present in ancient colinear blocks deviated from the expected number, if all genes were randomly distributed on the chromosomes. Chromosomes significantly deviating (Fisher's exact test and false discovery rate correction) are mentioned in the main text and are shown in File S3 and Figure S43. Genes detected to be derived from ancestor 7 and ancestor 12 karyotpyes can be traced to extant chromosomes (File S3).

### GO bias analyses and GO word cloud presentation

Analyses were conducted as described previously (Widiez *et al.*, 2014), using the GOstats R package and Fisher's exact test with fdr correction. Visualization of the GO terms was implemented using word clouds via the http://www.wordle.net application. The weight of the given terms was defined as the −log10(*q*-values) and the colour scheme used for the visualization was red for under-represented GO terms and green for those over-represented. Terms with stronger representation, i.e. weight >4, were represented with darker colours.

### Circos plots

For the integrative visualization of the individual genomic features a karyotype ideogram was created and tracks were plotted with CIRCOS v0.67-6 (Krzywinski *et al.*, 2009). For each feature track it is highlighted in the corresponding figure legend whether feature raw counts/values were used for visualization or if chromosomes were split into smaller windows (specifying the window size in kbp and window overlaps/jumps in kbp) using the counts/values window average for visualization. If indicated, feature counts/values window averages (cvwa) were normalized by scaling between a range of 0 and 1 per chromosome using the following equation:

$$\text{normalized window average}_{chr}(\text{cvwa}_{i_{chr}}) = \frac{\text{cvwa}_{i_{chr}} - \text{cvwa}_{chr_{min}}}{\text{cvwa}_{ch_{max}} - \text{cvwa}_{chr_{min}}}$$

For normalized comparison of embryophyte chromosome structure refer to Appendix S1, Supplementary Material, Section III; for phylostratigraphy analyses to Appendix S1, Supplementary Material, Section IV.

### Availability of data and material

The data reported in this paper are tabulated in Experimental Procedures and Supporting Information, are archived at the NCBI SRA and have been made available using the comparative genomics (CoGe) environment of CyVerse (cyverse.org) via https://genomevolution.org/coge/GenomeView.pl?gid=33928. Novel data presented with this study comprise Villersexel and Kaskaskia genomic DNA (SRX037761, SRX030894, SRP091316), genomic BAC end data (KS521087–KS697761), RNA-seq data (Table S6 and File S3 – available from phytozome.org), CAP-capture and BS-seq data (Table S10), and Goldengate SNP bead array data (File S3). See also section Availability of gene models and additional data.

Requests for materials should be addressed to stefan.rensing@biologie.uni-marburg.de.

## AUTHORS' CONTRIBUTIONS

AS, ADZ, ACC, AW, CVC, DL, FH, FMu, FMa, GB, HG, JP, JSa, JJ, GAT, JM, JF, JMC, KV, KKU, LEG, LS, MH, MT, MP, MvB, NvG, OS, PR, RM, RH, SNWH, SS, SAR, SFM, TW, WM, YK, YZ analysed data or performed experiments. AL, CR, DWS, ELD, FT, FWL, GW, JCVA, JG, PFP, SAR, SG, RR, RSQ, YZ contributed samples, materials or data. DSR, DG, JSc, JSa, GAT, JMC, KV, KFXM, RR, SAR supervised part of the research. ACC, DL, FMa, SAR wrote the paper with help by SG, KFXM, DWS and contributions by all authors. JSc and SAR coordinated the project.

## ACKNOWLEDGEMENTS

We thank Richard Haas, Faezeh Donges, Marco Göttig and Katrin Kumke for technical assistance. We thank Walter Sanseverino and Riccardo Aliese (Sequentia Biotech) for assistance in TE RNAseq analyses. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. Support to RR and SAR by the German Research Foundation (DFG RE 837/10-2), the Excellence Initiative of the German Federal and State Governments (EXC 294), and by the German Federal Ministry of Education and Research (BMBF FRISYS), is highly appreciated. CoGe is supported by the US National Science Foundation under Award Numbers IOS-339156 and IOS-1444490, CyVerse is supported by the U.S. National Science Foundation under Award Numbers DBI-0735191 and DBI-1265383. YK and ACC are grateful for support from the UK Biological Sciences and Biotechnology Research Council (Grant BB/F001797/1). KV acknowledges the Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks' Project (no 01MR0410W) of Ghent University. JC is grateful for support from the Spanish Ministerio de Economía y Competitividad (Grant AGL2013-43244-R). RSQ is grateful to Monsanto (St. Louis, MO, USA) for sequencing genomic DNA of *P. patens* accession Kaskaskia. The 1000 Plants (1 KP) initiative, led by GKSW, is funded by the Alberta Ministry of Innovation and Advanced Education, Alberta Innovates Technology Futures (AITF), Innovates Centres of Research Excellence (iCORE), Musea Ventures, BGI-Shenzhen and China National Genebank (CNGB). TW was supported by EMBO Long-Term Fellowships (ALTF 1166-2011) and by Marie Curie Actions (European Commission EMBOCOFUND2010, GA-2010-267146). The work conducted at PGSB was supported by the German Research Foundation (SFB924) and German Ministry of Education and Research (BMBF, 031A536/de.NBI). The authors declare that they have no competing interests.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

**Appendix S1.** Supplementary Materials I–IV, Experimental Procedures, and Results including Tables S1–S23, Figures S1–S50, and References.

**File S1.** v3.1 + v3.3 annotation.

**File S2.** Plots of markers, TE methylation and histone modification, phenotypic differences of *P. patens* accessions, sRNA northern blots.

**File S3.** Synteny analyses, JGI gene atlas samples, NCLDV clusters/genes, JGI bead array SNP QC.

## REFERENCES


**Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A. and Mason, C.E.** (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87.

**Arif, M.A., Fattash, I., Ma, Z., Cho, S.H., Beike, A.K., Reski, R., Axtell, M.J. and Frank, W.** (2012) DICER-LIKE3 activity in Physcomitrella patens DICER-LIKE4 mutants causes severe developmental dysfunction and sterility. *Mol. Plant*, **5**, 1281–1294.

**Bainard, J.D. and Newmaster, S.G.** (2010) Endopolyploidy in bryophytes: widespread in mosses and absent in liverworts. *J. Bot.* **2010**, 7.

**Bainard, J.D. and Villarreal, J.C.** (2013) Genome size increases in recently diverged hornwort clades. *Genome*, **56**, 431–435.

**Bainard, J.D., Forrest, L.L., Goffinet, B. and Newmaster, S.G.** (2013) Nuclear DNA content variation and evolution in liverworts. *Mol. Phylogenet. Evol.* **68**, 619–627.

**Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J. and Rieseberg, L.H.** (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* **25**, 2445–2455.

**Beike, A.K., von Stackelberg, M., Schallenberg-Rudinger, M., Hanke, S.T., Follo, M., Quandt, D., McDaniel, S.F., Reski, R., Tan, B.C. and Rensing, S.A.** (2014) Molecular evidence for convergent evolution and allopolyploid speciation within the Physcomitrium-Physcomitrella species complex. *BMC Evol. Biol.* **14**, 158.

**Benson, G.** (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.

**Bewick, A.J., Niederhuth, C.E., Ji, L., Rohr, N.A., Griffin, P.T., Leebens-Mack, J. and Schmitz, R.J.** (2017) The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biol.* **18**, 65.

**Blanc, G., Agarkova, I., Grimwood, J. *et al.*** (2012) The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* **13**, R39.

**Bolger, A.M., Lohse, M. and Usadel, B.** (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

**Cao, J., Schneeberger, K., Ossowski, S. *et al.*** (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963.

**Chakravarti, A., Lasher, L.K. and Reefer, J.E.** (1991) A maximum likelihood method for estimating genome length using genetic linkage data. *Genetics*, **128**, 175–182.

**Chaudhuri, P. and Marron, J.S.** (1999) SiZer for exploration of structures in curves. *J. Am. Stat. Assoc.* **94**, 807–823.

**Dangwal, M., Kapoor, S. and Kapoor, M.** (2014) The PpCMT chromomethylase affects cell growth and interacts with the homolog of LIKE HETEROCHROMATIN PROTEIN 1 in the moss *Physcomitrella patens*. *Plant J.* **77**, 589–603.

**De Bodt, S., Maere, S. and Van de Peer, Y.** (2005) Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.* **20**, 591–597.

**Dellaporta, S.L., Wood, J. and Hicks, J.B.** (1983) A plant DNA minipreparation: Version II. *Plant Mol. Biol. Rep.* **1**, 19–21.

**Devos, N., Szovenyi, P., Weston, D.J., Rothfels, C.J., Johnson, M.G. and Shaw, A.J.** (2016) Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnopsida (Bryophyta). *New Phytol.* **211**, 300–318.

**Dolgin, E.S. and Charlesworth, B.** (2008) The effects of recombination rate on the distribution and abundance of transposable elements. *Genetics*, **178**, 2169–2177.

**Ellinghaus, D., Kurtz, S. and Willhoeft, U.** (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.

**Feng, S., Cokus, S.J., Zhang, X. *et al.*** (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA*, **107**, 8689–8694.

**Feng, J., Liu, T., Qin, B., Zhang, Y. and Liu, X.S.** (2012) Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740.

**Finn, R.D., Coggill, P., Eberhardt, R.Y. *et al.*** (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285.

**Fishman, L., Kelly, A.J., Morgan, E. and Willis, J.H.** (2001) A genetic map in the Mimulus guttatus species complex reveals transmission ratio distortion due to heterospecific interactions. *Genetics*, **159**, 1701–1716.

**Flutre, T., Duprat, E., Feuillet, C. and Quesneville, H.** (2011) Considering transposable element diversification in de novo annotation approaches. *PLoS ONE*, **6**, e16526.

**Fuchs, J., Demidov, D., Houben, A. and Schubert, I.** (2006) Chromosomal histone modification patterns – from conservation to diversity. *Trends Plant Sci.* **11**, 199–208.

**Foissac, S., Bardou, P., Moisan, A., Cros, M.J. and Schiex, T.** (2003) EUGENE'HOM: A generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Res.* **31**, 3742–3745.

**Gernand, D., Demidov, D. and Houben, A.** (2003) The temporal and spatial pattern of histone H3 phosphorylation at serine 28 and serine 10 is similar in plants but differs between mono- and polycentric chromosomes. *Cytogenet. Genome Res.* **101**, 172–176.

**Grabherr, M.G., Haas, B.J., Yassour, M. *et al.*** (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.

**Gremme, G., Brendel, V., Sparks, M.E. and Kurtz, S.** (2005) Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978.

**Harrison, C.J., Roeder, A.H., Meyerowitz, E.M. and Langdale, J.A.** (2009) Local cues and asymmetric cell divisions underpin body plan transitions in the moss *Physcomitrella patens*. *Curr. Biol.* **18**, 18.

**Hartmann, M.A.** (1998) Plant sterols and the membrane environment. *Trends Plant Sci.* **3**, 170–175.

**Hiss, M., Meyberg, R., Westermann, J., Haas, F.B., Schneider, L., Schallenberg-Rudinger, M., Ullrich, K.K. and Rensing, S.A.** (2017) Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute. *Plant J.* **90**, 606–620 https://doi.org/10.1111/tpj.13501.

**Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V. and Quesneville, H.** (2014) PASTEC: an automatic transposable element classification tool. *PLoS ONE*, **9**, e91929.

**Horst, N.A., Katz, A., Pereman, I., Decker, E.L., Ohad, N. and Reski, R.** (2016) A single homeobox gene triggers phase transition, embryogenesis and asexual reproduction. *Nat. Plants*, **2**, 15209.

**Hu, R., Xiao, L., Bao, F., Li, X. and He, Y.** (2016) Dehydration-responsive features of Atrichum undulatum. *J. Plant. Res.* **129**, 945–954.

**Hutter, S., Vilella, A.J. and Rozas, J.** (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, **7**, 409.

**Ibarra, C.A., Feng, X., Schoft, V.K. *et al.*** (2012) Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science*, **337**, 1360–1364.

**Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C. and Lander, E.S.** (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96.

**Kaessmann, H.** (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326.

**Kamisugi, Y., von Stackelberg, M., Lang, D., Care, M., Reski, R., Rensing, S.A. and Cuming, A.C.** (2008) A sequence-anchored genetic linkage map for the moss, *Physcomitrella patens*. *Plant J.* **56**, 855–866.

**Kawashima, T. and Berger, F.** (2014) Epigenetic reprogramming in plant sexual reproduction. *Nat. Rev. Genet.* **15**, 613–624.

**Keibler, E. and Brent, M.R.** (2003) Eval: a software package for analysis of genome annotations. *BMC Bioinformatics*, **4**, 50.

**Kent, W.J.** (2002) BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.

**Khraiwesh, B., Arif, M.A., Seumel, G.I., Ossowski, S., Weigel, D., Reski, R. and Frank, W.** (2010) Transcriptional control of gene expression by microRNAs. *Cell*, **140**, 111–122.
**Krueger, F. and Andrews, S.R.** (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
**Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A.** (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.
**Lamb, J.C., Yu, W., Han, F. and Birchler, J.A.** (2007) Plant chromosomes from end to end: telomeres, heterochromatin and centromeres. *Curr. Opin. Plant Biol.* **10**, 116–122.
**Lang, D., Weiche, B., Timmerhaus, G., Richardt, S., Riano-Pachon, D.M., Correa, L.G., Reski, R., Mueller-Roeber, B. and Rensing, S.A.** (2010) Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.* **2**, 488–503.
**Li, H. and Durbin, R.** (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
**Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.; Genome Project Data Processing Subgroup** (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
**Li, Y.H., Zhou, G., Ma, J. *et al.*** (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052.
**Lowe, T.M. and Eddy, S.R.** (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964.
**Martinez, G. and Slotkin, R.K.** (2012) Developmental relaxation of transposable element silencing in plants: functional or byproduct? *Curr. Opin. Plant Biol.* **15**, 496–502.
**Maumus, F., Epert, A., Nogue, F. and Blanc, G.** (2014) Plant genomes enclose footprints of past infections by giant virus relatives. *Nat. Commun.* **5**, 4268.
**McDaniel, S.F., von Stackelberg, M., Richardt, S., Quatrano, R.S., Reski, R. and Rensing, S.A.** (2010) The speciation history of the Physcomitrium-Physcomitrella species complex. *Evolution*, **64**, 217–231.
**McKenna, A., Hanna, M., Banks, E. *et al.*** (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.
**Melters, D.P., Bradnam, K.R., Young, H.A. *et al.*** (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10.
**Newton, A.E., Wikström, N., Bell, N., Forrest, L.L. and Ignatov, M.S.** (2006) Dating the diversification of the pleurocarpous mosses. In *Pleurocarpous mosses: Systematics and Evolution*. (Tangney, N., ed). Boca Raton: CRC Press, Systematics Association, pp. 329–358.
**Niederhuth, C.E., Bewick, A.J., Ji, L. *et al.*** (2016) Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194.
**Oliver, M.J., Dowd, S.E., Zaragoza, J., Mauget, S.A. and Payton, P.R.** (2004) The rehydration transcriptome of the desiccation-tolerant bryophyte *Tortula ruralis*: transcript classification and analysis. *BMC Genom.* **5**, 89.
**Olsen, J.L., Rouze, P., Verhelst, B. *et al.*** (2016) The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature*, **530**, 331–335.
**Ortiz-Ramirez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., Feijo, J.A. and Becker, J.D.** (2015) A transcriptome atlas of *Physcomitrella patens* provides insights into the evolution and development of land plants. *Mol. Plant*, **9**, 205–220.
**Perroud, P.F., Cove, D.J., Quatrano, R.S. and McDaniel, S.F.** (2011) An experimental method to facilitate the identification of hybrid sporophytes in the moss *Physcomitrella patens* using fluorescent tagged lines. *New Phytol.* **2**, 1469–8137.
**Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y. and Vandepoele, K.** (2012) i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11.
**Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M. and Anxolabehere, D.** (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* **1**, 166–175.
**Rensing, S.A.** (2014) Gene duplication as a driver of plant morphogenetic evolution. *Curr. Opin. Plant Biol.* **17C**, 43–48.
**Rensing, S.A., Ick, J., Fawcett, J.A., Lang, D., Zimmer, A., Van de Peer, Y. and Reski, R.** (2007) An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol. Biol.* **7**, 130.
**Rensing, S.A., Lang, D., Zimmer, A.D. *et al.*** (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.
**Rensing, S.A., Beike, A.K. and Lang, D.** (2012) Evolutionary importance of generative polyploidy for genome evolution of haploid-dominant land plants. In *Plant Genome Diversity* (Greilhuber, J., Wendel, J.F., Leitch, I.J. and Doležel, J., eds). Vienna, New York: Springer, pp. 295–305.
**Rensing, S.A., Sheerin, D.J. and Hiltbrunner, A.** (2016) Phytochromes: more than meets the eye. *Trends Plant Sci.* **21**, 543–546.
**Reski, R., Faust, M., Wang, X.H., Wehe, M. and Abel, W.O.** (1994) Genome analysis of the moss *Physcomitrella patens* (Hedw.) B.S.G. *Mol. Gen. Genet.* **244**, 352–359.
**Sakakibara, K., Ando, S., Yip, H.K., Tamada, Y., Hiwatashi, Y., Murata, T., Deguchi, H., Hasebe, M. and Bowman, J.L.** (2013) KNOX2 genes regulate the haploid-to-diploid morphological transition in land plants. *Science*, **339**, 1067–1070.
**Salse, J.** (2012) In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr. Opin. Plant Biol.* **15**, 122–130.
**Salse, J., Abrouk, M., Murat, F., Quraishi, U.M. and Feuillet, C.** (2009) Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief. Bioinform.* **10**, 619–630.
**Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J. and Shoemaker, R.C.** (2004) Mining EST databases to resolve evolutionary events in major crop species. *Genome*, **47**, 868–876.
**Schween, G., Gorr, G., Hohe, A. and Reski, R.** (2003) Unique tissue-specific cell cycle in *Physcomitrella*. *Plant Biol.* **5**, 50–58.
**Schween, G., Egener, T., Fritzkowsky, D. *et al.*** (2005) Large-scale analysis of 73,329 gene-disrupted *Physcomitrella* mutants: production parameters and mutant phenotypes. *Plant Biol.* **7**, 238–250.
**Simillion, C., Janssens, K., Sterck, L. and Van de Peer, Y.** (2008) i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics*, **24**, 127–128.
**Smit, A.F.A., Hubley, R. and Green, P.** (1996) RepeatMasker Open-3.0. URL http://www.repeatmasker.org.(unpublished), 2004.
**Spannagl, M., Nussbaumer, T., Bader, K.C., Martis, M.M., Seidel, M., Kugler, K.G., Gundlach, H. and Mayer, K.F.** (2016) PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Res.* **44**, D1141–D1147.
**Stroud, H., Otero, S., Desvoyes, B., Ramirez-Parra, E., Jacobsen, S.E. and Gutierrez, C.** (2012) Genome-wide analysis of histone H3.1 and H3.3 variants in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **109**, 5370–5375.
**Szovenyi, P., Ricca, M., Hock, Z., Shaw, J.A., Shimizu, K.K. and Wagner, A.** (2013) Selection is no more efficient in haploid than in diploid life stages of an angiosperm and a moss. *Mol. Biol. Evol.* **30**, 1929–1939.
**Szovenyi, P., Perroud, P.F., Symeonidi, A., Stevenson, S., Quatrano, R.S., Rensing, S.A., Cuming, A.C. and McDaniel, S.F.** (2014) De novo assembly and comparative analysis of the *Ceratodon purpureus* transcriptome. *Mol. Ecol. Resour.* **15**, 203–215.
**Trapnell, C., Pachter, L. and Salzberg, S.L.** (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
**Van de Peer, Y., Mizrachi, E. and Marchal, K.** (2017) The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424.
**Van de Velde, J., Van Bel, M., Van Eechoutte, D. and Vandepoele, K.** (2016) A collection of conserved non-coding sequences to study gene regulation in flowering plants. *Plant Physiol.* **171**, 2586–2598.
**Vanneste, K., Baele, G., Maere, S. and Van de Peer, Y.** (2014) Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res.* **24**, 1334–1347.
**Vanneste, K., Sterck, L., Myburg, A.A., Van de Peer, Y. and Mizrachi, E.** (2015) Horsetails are ancient polyploids: evidence from *Equisetum giganteum*. *Plant Cell*, **27**, 1567–1578.

**Vekemans, D., Proost, S., Vanneste, K., Coenen, H., Viaene, T., Ruelens, P., Maere, S., Van de Peer, Y. and Geuten, K.** (2012) Gamma paleo-hexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Mol. Biol. Evol.* **29**, 3793–3806.

**Vives, C., Charlot, F., Mhiri, C., Contreras, B., Daniel, J., Epert, A., Voytas, D.F., Grandbastien, M.A., Nogue, F. and Casacuberta, J.M.** (2016) Highly efficient gene tagging in the bryophyte *Physcomitrella patens* using the tobacco (*Nicotiana tabacum*) Tnt1 retrotransposon. *New Phytol.* **212**, 759–769.

**Wang, G., Zhang, X. and Jin, W.** (2009) An overview of plant centromeres. *J. Genet. Genomics* **36**, 529–537.

**Widiez, T., Symeonidi, A., Luo, C., Lam, E., Lawton, M. and Rensing, S.A.** (2014) The chromatin landscape of the moss *Physcomitrella patens* and its dynamics during development and drought stress. *Plant J.* **79**, 67–81.

**Wright, S.I., Agrawal, N. and Bureau, T.E.** (2003) Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Res.* **13**, 1897–1903.

**Wu, T.D., Reeder, J., Lawrence, M., Becker, G. and Brauer, M.J.** (2016) GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol. Biol.* **1418**, 283–334.

**Zemach, A., McDaniel, I.E., Silva, P. and Zilberman, D.** (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, **328**, 916–919.

**Zhang, Y., Liu, T., Meyer, C.A. *et al.*** (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137.

**Zilberman, D.** (2017) An evolutionary case for functional gene body methylation in plants and animals. *Genome Biol.* **18**, 87.

**Zimmer, A.D., Lang, D., Buchta, K., Rombauts, S., Nishiyama, T., Hasebe, M., Van de Peer, Y., Rensing, S.A. and Reski, R.** (2013) Reannotation and extended community resources for the genome of the non-seed plant *Physcomitrella patens* provide insights into the evolution of plant gene structures and functions. *BMC Genom.* **14**, 498.

## 6.4 Publication 4

In my fourth publication I could show that the microarray expression data from the Gransden ecotype can be used to predict expression strength across ecotypes. We tested an endogenous promoter of the *lhcsr1* gene which showed a high expression across most developmental stages in the Gransden ecotype microarray data. In Reute we also find a high expression in most of the tissues. These experiments also showed that Reute can be as easily transformed as Gransden. As part of this publication I developed a plate reader measurement system that allows fluorescence measurement of living protoplasts, which in our case contained transiently transformed promoter-reporter gene constructs. With this system I not only compared the expression strength of the two commonly used promoters 2x 35S and the rice actin1 with the endogenous lhcsr1 promoter but also tested *lhcsr1* promoter fragments for their activity as compared to the full length *lhcsr1* promoter. Additionally we analysed the codon usage bias in *P. patens* and added GFP variants with adapted codons to our analysis. I show that a higher expression level in *P. patens* can be achieved by codon optimization and thereby provide a GFP variant showing higher *in vivo* signal intensities. These results will be especially useful for biotechnological applications where high expression levels are important to achieve high protein yields.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 31 October 2017
doi: 10.3389/fpls.2017.01842



OPEN ACCESS

**Edited by:**
*Henrik Toft Simonsen, Technical University of Denmark, Denmark*

**Reviewed by:**
*Kashmir Singh, Panjab University, India*
*Fabien Nogué, INRA Centre Versailles-Grignon, France*

***Correspondence:**
*Stefan A. Rensing*
*stefan.rensing@biologie.uni-marburg.de*

**†Present address:**
*Lucas Schneider, Institute for Transfusion Medicine and Immunohematology, Johann Wolfgang Goethe University, German Red Cross Blood Donor Service, Frankfurt, Germany*
*Aikaterini Symeonidi, Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain*
*Kristian K. Ullrich, Max Planck Institute for Evolutionary Biology, Plön, Germany*
*Mareike Schallenberg-Rüdinger, Institut für Zelluläre und Molekulare Botanik, Abteilung Molekulare Evolution, Universität Bonn, Bonn, Germany*

**Specialty section:**
*This article was submitted to Plant Biotechnology, a section of the journal Frontiers in Plant Science*

**Received:** *19 August 2017*
**Accepted:** *10 October 2017*
**Published:** *31 October 2017*

**Citation:**
*Hiss M, Schneider L, Grosche C, Barth MA, Neu C, Symeonidi A, Ullrich KK, Perroud P-F, Schallenberg-Rüdinger M and Rensing SA (2017) Combination of the Endogenous lhcsr1 Promoter and Codon Usage Optimization Boosts Protein Expression in the Moss Physcomitrella patens. Front. Plant Sci. 8:1842. doi: 10.3389/fpls.2017.01842*

# Combination of the Endogenous *lhcsr1* Promoter and Codon Usage Optimization Boosts Protein Expression in the Moss *Physcomitrella patens*

***Manuel Hiss*[1], *Lucas Schneider*[1†], *Christopher Grosche*[1], *Melanie A. Barth*[1], *Christina Neu*[1], *Aikaterini Symeonidi*[1†], *Kristian K. Ullrich*[1†], *Pierre-François Perroud*[1], *Mareike Schallenberg-Rüdinger*[1†] *and Stefan A. Rensing*[1,2*]**

[1] *Plant Cell Biology, Faculty of Biology, University of Marburg, Marburg, Germany,* [2] *BIOSS Centre for Biological Signaling Studies, University of Freiburg, Freiburg im Breisgau, Germany*

The moss *Physcomitrella patens* is used both as an evo-devo model and biotechnological production system for metabolites and pharmaceuticals. Strong *in vivo* expression of genes of interest is important for production of recombinant proteins, e.g., selectable markers, fluorescent proteins, or enzymes. In this regard, the choice of the promoter sequence as well as codon usage optimization are two important inside factors to consider in order to obtain optimum protein accumulation level. To reliably quantify fluorescence, we transfected protoplasts with promoter:GFP fusion constructs and measured fluorescence intensity of living protoplasts in a plate reader system. We used the red fluorescent protein mCherry under 2x *35S* promoter control as second reporter to normalize for different transfection efficiencies. We derived a novel endogenous promoter and compared deletion variants with exogenous promoters. We used different codon-adapted green fluorescent protein (GFP) genes to evaluate the influence of promoter choice and codon optimization on protein accumulation in *P. patens*, and show that the promoter of the gene of *P. patens* chlorophyll a/b binding protein *lhcsr1* drives expression of GFP in protoplasts significantly (more than twofold) better than the commonly used 2x *35S* promoter or the rice *actin1* promoter. We identified a shortened 677 bp version of the *lhcsr1* promoter that retains full activity in protoplasts. The codon optimized GFP yields significantly (more than twofold) stronger fluorescence signals and thus demonstrates that adjusting codon usage in *P. patens* can increase expression strength. In combination, new promotor and codon optimized GFP conveyed sixfold increased fluorescence signal.

**Keywords: *Physcomitrella patens*, codon usage, chlorophyll a/b binding protein, promoter, codon bias, green fluorescent protein (GFP), *lhcsr1*, fluorescence normalization**

## INTRODUCTION

The strength of protein expression can be influenced by many factors including outside factors such as culture conditions, or inside factors as codon usage or transcription/translation system (Ullrich et al., 2015). Both promoter sequence and coding sequence can be optimized to improve final protein accumulation. The constructs of the first stable mutant lines in *Physcomitrella patens* contained resistance cassettes controlled by the cauliflower mosaic virus (CaMV) *19S* and *35S* promoters and the *Rhizobium radobacter* (previously *Agrobacterium tumefaciens*) nopaline synthase gene (*nos*) promoter (Schaefer et al., 1991). As *P. patens* was further developed as a plant model, classical strong angiosperm promoters such as the *Oryza sativa* (rice) *actin1* promoter (McElroy et al., 1990) or the *Zea mays* (maize) *ubiquitin* promoter (Christensen et al., 1992) were used successfully to drive protein accumulation in moss (Bezanilla et al., 2003; Horstmann et al., 2004). Inducible expression systems have also been established in *P. patens*, such as the beta-estradiol inducible one (Kubo et al., 2013) or the induction by elevated temperature using a *Glycine max* (soybean) heat shock protein promoter (Saidi et al., 2005). In a previous study, the activity of several promoters was studied by transient *P. patens* protoplast transfection of promoter:luciferase fusion constructs; here, the *actin1* promoter showed 10 times the expression level of the single CaMV *35S* promoter and 1.6 times the level of the 2x CaMV *35S* promoter (Horstmann et al., 2004). The same study also included endogenous 5′ sequences of the genes α1,3-fucosyltransferase and β1,2-xylosyltransferase (*fuc-t, xyl-t*) that were further characterized *via* deletion constructs, with the 5′-*fuc-t* showing almost double activity as compared to the single CaMV *35S* promoter. To confer strong expression, other endogenous promoters were used, e.g., different *tubulin* (Jost et al., 2005) or *actin* promoters (Weise et al., 2006), all of which showed stronger expression than the CaMV *35S* promoter, that was later shown to yield only mediocre expression in *P. patens*, especially in the dark (Saidi et al., 2009).

Chlorophyll a/b binding (CAB) proteins are part of the light harvesting complex (LHC) of photosynthetic eukaryotes. *Cab* promoter sequences have been used as strong endogenous and exogenous promoters, e.g., in the charophyte alga *Closterium peracerosum–strigosum–littorale* complex, where an endogenous promoter of a *cab* gene has been used to drive expression of fluorescence marker genes (Abe et al., 2008). The rice *cab1R* gene promoter was used for transient expression of β-glucuronidase (*gus*) in *Nicotiana tabacum*, *Z. mays* and *O. sativa* leaves (Luan and Bogorad, 1992). In addition to the common LHC gene set, a LHC-like protein called LHCSR (or Li818) is present in association with LHC in phylogenetically diverse algae as *Chlamydomonas reinhardtii* (Li et al., 2000) and *Ectocarpus siliculosus*, but not in seed plants (Koziol et al., 2007; Dittami et al., 2010). LHC-like protein expression is regulated by light and stress conditions. In *P. patens* two *lhcsr* gene copies have been identified (Gerotto et al., 2011). *lhcsr1* is induced by high light (450 $\mu mol/m^{-2}$ $s^{-2}$), whereas *lhcsr2* is expressed in low temperature and low light conditions (Gerotto et al., 2011). Together with the protein PSBS, the LHCSR proteins are responsible for the non-photochemical quenching in *P. patens* (Alboresi et al., 2010). The mechanisms of photoprotection of *lhcsr1 via* two dissipative states has recently been revealed (Kondo et al., 2017), as well as its modulation *via* zeaxanthin binding and low pH (Pinnola et al., 2017).

A second important inside factor influencing protein production is the codon usage in the RNA sequence (Quax et al., 2015). Due to the redundancy of the genetic triplet code, almost all amino acids are encoded by more than one codon. Which codon is used to which degree depends, e.g., on the species and on availability of tRNAs (Komar, 2016). Codon usage can be influenced by mutations and affects expression speed or accuracy. When trying to express a gene sequence from one species in another species, the codon usage often needs to be adjusted to fit the target species' codon frequencies. For example, the codon usage of the GFP, that is used in many organisms, e.g., to localize proteins by tagging, has been optimized for different organisms like *Saccharomyces cerevisiae* (yeast) (Cormack et al., 1997) or the alga *C. reinhardtii* (Fuhrmann et al., 1999), leading to stronger GFP signals. For plants a soluble modified GFP (smGFP) was created by site-directed mutagenesis that shows stronger GFP accumulation in *Arabidopsis thaliana* and therefore a stronger signal than the wild-type GFP (Castillo-Davis et al., 2002). Codon usage can vary not only between species but also within a species. In a subgroup of genes from one species, a bias for certain codons can be found, e.g., in highly expressed genes of *Caenorhabditis elegans*, *Drosophila melanogaster*, and *A. thaliana* codon usage differs from more weakly expressed genes of the same species (Duret and Mouchiroud, 1999). In *P. patens*, this codon usage bias seems to be driven by a combination of weak natural selection and the predominant mutational biases (Szovenyi et al., 2017).

Here we used the endogenous promoter sequence of *lhcsr1* to drive gene expression of the GFP in *P. patens* and compared it to the double CaMV *35S* (2x p*35S*) and the rice *actin1* promoter (McElroy et al., 1990). In parallel, we designed two GFP versions with different codon usage and evaluated their GFP accumulation to the smGFP using a novel bimodal fluorescence readout system that allows to normalize the signal of interest in order to account for fluctuations in transfection efficiency. We find that the *lhcsr1* gene promoter increases signal intensity 1.7-fold as compared to the 2x *35S* promoter. In combination with the codon-optimized GFP the signal even increases 5.7-fold as compared to the 2x *35S* promoter.

## RESULTS

### Fluorescence Readout System

We designed a novel bimodal readout system that can be used to measure reporter protein fluorescence *in vivo*. We opted to use a microplate reader to measure fluorescence to allow high throughput measurement. For all fluorescence measurements, we used living protoplasts in transfection regeneration medium (Hohe et al., 2004). Initial tests showed that a minimum number of 2,000 protoplasts are necessary to get a signal above background fluorescence (Supplementary Figure 1).

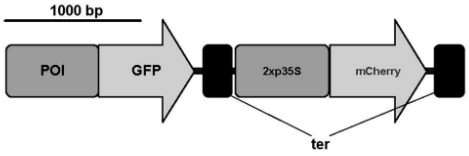



**FIGURE 1 |** Constructs used for transient expression of GFP and mCherry. Promoter of interest (POI); green fluorescent protein sequence (GFP); 2x *35S* promoter (2x p*35S*); *35S* terminator (ter). Scale bar represents 1,000 base pairs. A modified pAM-PAT vector (accession number: AY436765) was used as backbone for transient constructs. Figure was prepared with IBS (Liu et al., 2015).

The constructs used for transient transfection contained the promoter:GFP fusion and additionally a normalization cassette consisting of the 2x *35S* promoter, the mCherry gene sequence and a *35S* terminator (**Figure 1**). We selected the 2x *35S* promoter because it is widely used in the *P. patens* community and shows a medium gene expression in *P. patens*, which makes it a good candidate for comparisons with new promoters. The fluorescence from this second reporter was used to normalize the GFP fluorescence values, which show high between-experiment variation due to different transformation efficiencies (Supplementary Figure 2 and Supplementary Table 1). By integrating 2xp*35S*:mCherry into the same plasmid as the GFP fusion of interest, we also account for uptake of multiple plasmids during transformation. To test the feasibility of our plate reader measurement system, we performed protoplastation of protonemal tissue from stable mutant lines that express GFP or mCherry (Perroud et al., 2011), mixed the protoplasts and found that we can measure both signals in parallel from the same well. Additionally, we prepared a dilution series and found a linear relationship between number of protoplasts and fluorescence intensity for both reporters (Supplementary Figure 3).

## Highly Expressed Genes Show Codon Usage Bias in *P. patens*

We calculated the codon frequencies for all v1.6 genes of *P. patens* and grouped them by expression strength based on a broad range of transcriptome microarray experiments (Hiss et al., 2014). The top 233 highly expressed genes (0.9%) are enriched in the Gene Ontology (GO) terms translation, gene expression, and protein synthesis (Supplementary Figure 4). By comparing the codon usage of these highly expressed genes with the codon usage of all genes, we find a significant codon usage bias (Fisher's Exact Test, FDR adjusted $p < 0.05$) for nine amino acids, namely Glu, Phe, Tyr, Cys, His, Gln, Ile, Asn, and Lys (**Figure 2A** and Supplementary File 1). These codons are preferred not only in *P. patens* highly expressed genes but also in other organisms [**Figure 2B**; *Arabidospsis thaliana* (Wright et al., 2004; Morton and Wright, 2007), *Saccharomyces cerevisiae* (dos Reis and Wernisch, 2009), *Schizosaccharomyces pombe* (Hiraoka et al., 2009), and *Homo sapiens* (van Hemert and Berkhout, 1995)]. Using microarray data (Hiss et al., 2014), we can confidently detect the first biased codon usage (Lys) already within the top 1,967 highly expressed genes (7.3% of all genes measured by the array), and the codon usage bias for all nine amino acids can be detected based on the top 291 highly expressed genes (1.1%, Supplementary File 2). Thus, by using the biased codon usage as an indicator, we detect around 7% of the *P. patens* genes as highly expressed. These genes also show high expression in microarray-based transcriptome studies [Supplementary Figure 5 (Hiss et al., 2014; Ortiz-Ramirez et al., 2015)]. For the highly expressed genes, we detect a GC bias toward higher GC content and a lower effective number of codons (ENC) as compared to the overall gene set (Supplementary Figure 6).

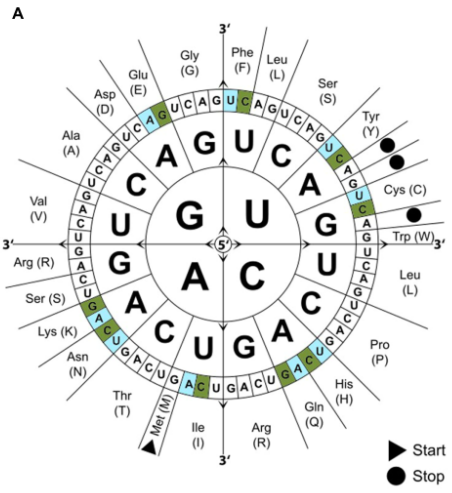
To determine whether the codon bias we have seen in highly expressed genes in *P. patens* can be used to enhance protein expression, we prepared GFP variants with adapted codon usage. Starting from the smGFP (Davis and Vierstra, 1998), we changed all triplets either to the one preferred in highly expressed genes, or to the one not preferred. This led to 39 changes in the coding sequence for "GFPhigh" and 58 changes for "GFPlow," respectively (Supplementary Figure 7 and Supplementary Table 2). The GC content of the smGFP was 43.7% as compared to 49.1% (GFPhigh) and 35.6% (GFPlow). In both adapted GFP versions, the resulting amino acid sequence was not changed.

All three GFP versions were combined with different promoters 2xp*35S*, p*actin1*, and p*lhcsr1* (1,956 bp upstream region of the *Physcomitrella patens lhcsr1* gene) and the mCherry normalization cassette. These constructs were transfected transiently into *P. patens* protoplasts and the fluorescence signal of GFP and mCherry measured *in vivo* with a microplate reader system.

The measurements were normalized and the 2x *35S* promoter in combination with the smGFP signal set to one. The rice *actin1* promoter:GFPhigh fusion is at 1.0 ± 0.6-fold in our system and therefore not significantly different from the 2x *35S* promoter with the same GFP version ($p = 0.15$, Student's *t*-test). We see a shift in signal intensity for the 2x *35S* promoter constructs with different GFP versions, with the GFPhigh giving the strongest signal at 1.8 ± 0.5-fold and the GFPlow the weakest at 0.4 ± 0.3-fold. Different signal intensities can also be seen with the different p*lhcsr1*:GFP combinations starting with GFPlow at 0.9 ± 0.8-fold to smGFP at 1.7 ± 0.3-fold up to GFPhigh at 5.7 ± 0.4-fold (**Figure 3**).

## *lhcsr1* Promoter Shows Strong Expression

We selected a strong and constitutively expressed gene from our transcriptome microarray data which cover several developmental stages and perturbations (Hiss et al., 2014). To find a suitable endogenous promoter we calculated the coefficient of variation (c. v.) of the expression values of all genes represented on the Combimatrix microarray (Wolf et al., 2010) and selected genes with high expression values and low coefficient of variation, leading to candidate genes with a high expression across many developmental stages and perturbations. Additionally we filtered for genes whose 3 kbp upstream region does not overlap with




B

| Amino acid | Codon | P. patens | A. thaliana | S. cerevisiae | S. pombe | H. Sapiens |
|---|---|---|---|---|---|---|
| Cys | UGC | + | + | - | + | + |
| | UGU | - | - | + | - | - |
| Glu | GAA | - | - | + | - | - |
| | GAG | + | + | - | + | + |
| Phe | UUC | + | + | + | + | + |
| | UUU | - | - | - | - | - |
| His | CAC | + | + | + | + | + |
| | CAU | - | - | - | - | - |
| Ile | AUA | - | - | / | - | - |
| | AUC | + | + | / | - | + |
| | AUU | | | | | |
| Lys | AAA | - | - | - | - | - |
| | AAG | + | + | + | + | + |
| Asn | AAC | + | + | + | + | + |
| | AAU | - | - | - | - | - |
| Gln | CAA | - | - | - | + | - |
| | CAG | + | + | + | - | + |
| Tyr | UAC | + | + | + | + | + |
| | UAU | - | - | - | - | - |

**FIGURE 2 |** Codon usage bias in *P. patens* as compared to other selected organisms. Codon usage differs between all genes and highly expressed genes. Codons significantly over represented in highly expressed genes are colored dark green/marked by a (+) and significantly under represented codons are colored light blue/marked by a (–) (*p*-adjusted $< 0.05$, Fisher's exact test). **(A)** Codon sun adapted from Wikimedia Commons with biased codons marked in green (more abundant) or blue (less abundant). **(B)** Table of amino acids, codons, and codon bias in different species. Codons without bias are marked (/).

another gene or transposable element (TE). The 3 kbp gene-free upstream region was chosen in order to prevent the regulatory sequences of the gene of interest to fall into the sequence of another gene, and in order not to select atypically short genes (Zimmer et al., 2013). The 3kbp upstream region further should not contain TEs because they are known to be silenced *via* methylation of the corresponding genomic region (Zemach et al., 2010; Widiez et al., 2014). We chose the cab protein LHCSR1 (Phypa_169593, Pp1s213_80V6.1, Pp3c9_3440V3.1) since the promoters of *cab* genes of plants have successfully been used as strong endogenous and exogenous promoters before (Abe et al., 2008). We used about 2,000 bp upstream of the coding sequence (Chr09:1,975,316. . .1,977,272) as putative promoter for cloning and subsequent transfection (expression profile of the *lhcsr1* gene, see Supplementary Figure 8). The *lhcsr1* gene does not show the bias in codon usage we generally observe in the top highly expressed genes (Supplementary Table 3). The constructs containing the full p*lhcsr1* promoter show higher signals than the corresponding 2x p*35S* constructs (**Figure 3**). We prepared eight shortened promoter sequences by restriction digestion of the original 1,956 bp sequence (**Figure 4**, constructs A–H). All promoter versions were fused to GFPhigh:*35S*-terminator and the mCherry normalization cassette, transfected transiently into moss protoplasts and the GFP and mCherry signal measured *in vivo* with the microplate reader system.

Based on restriction enzyme sites, we divided the 1,956 bp promoter sequence into four regions (**Figure 4**). Region I contains the 202 bp 5′-UTR and 84 bp of the upstream region. Region II consists of a 391-bp fragment, Region III of a 520-bp fragment, and Region IV of the 759 bp at the 5′ end of the chosen promoter sequence. The removal of the regions IV and III without modifying the rest of the promoter does not strongly affect promoter activity, and the remaining fragments display an activity between 121 and 129% compared to the full length (1,956 bp) promoter (**Figure 4**, constructs A, B, and C). None of the other partial deletions abolish totally the promoter activity, most of the regions I and II deletions display 12–20% total p*lhcsr1* value (**Figure 4**, constructs D–H). The only exception is the construct D that does not contain the activating region II, but still retains 35% promoter activity (**Figure 4**, construct D). The region IV appears to contain a domain able to activate transcription once it is fused to the 5′-UTR. This is especially interesting in contrast to construct E which has longer fragments for regions I and IV but a lower signal intensity.

We searched the potential promoter sequence for motifs using Signal Scan and the plant *cis*-acting regulatory DNA elements database [PLACE, (Higo et al., 1999)] as well as the Plant *cis*-acting regulatory element (PlantCARE) database. PLACE finds several known *cis*-acting elements that are associated with light-regulated genes, e.g., GT1-sites (Green et al., 1988) or I-boxes (Giuliano et al., 1988), whereas many TATA boxes are found in region IV, further away from the ATG but not in the 5′-UTR or the regions II and III (**Figure 5A**). PlantCARE predicts a TATA box core promoter element for both *lhcsr* genes within the 5′-UTR at around −130 bp of the translation start (**Figure 5B**). The transcription start site

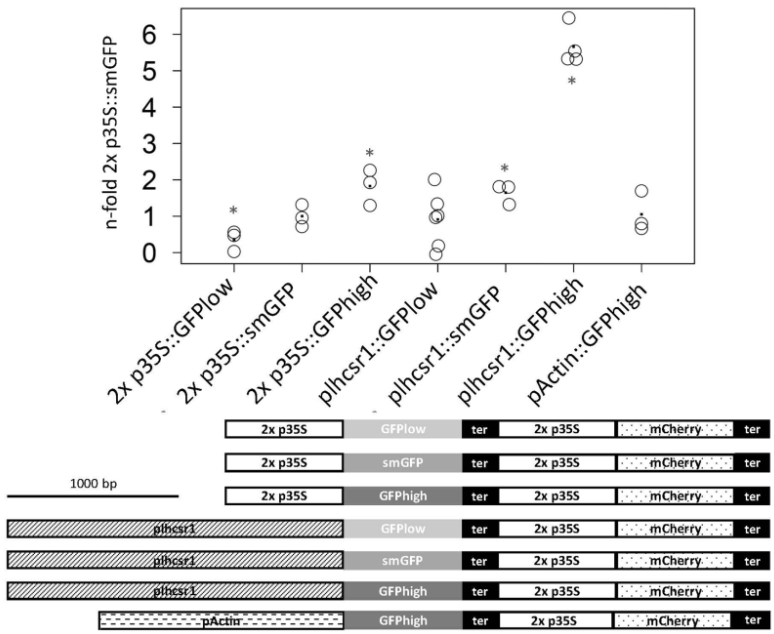



**FIGURE 3 |** Measurement of normalized GFP signal of seven different promoter:GFP fusion constructs. GFP and mCherry signal was measured with a microplate reader in transiently transfected protoplasts. GFP signal intensity is normalized by mCherry signal intensity and the 2x p*35S*:smGFP was set to one. Average value is marked by a dot (3–6 replicates). Asterisks mark significant difference from 2x p*35S*:smGFP (*$p < 0.05$, Student's *t*-test). The seven fusion constructs are depicted below the graph. Scale bar represents 1,000 base pairs. 2x *35S* promoter (2x p*35S*), *lhcsr1* promoter (p*lhcsr1*), rice actin 1 promoter (p*Actin*), soluble modified GFP (smGFP), codon adapted GFP (GFPlow, GFPhigh), and *35S* terminator (ter).

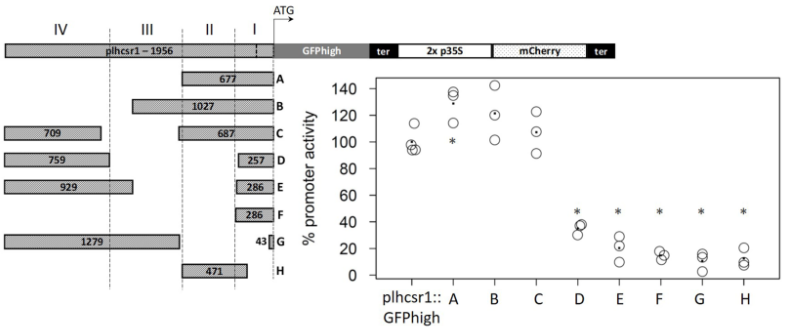



**FIGURE 4 |** Comparison of percentage promoter activity of nine different promoter fragments in GFPhigh fusion constructs. GFPhigh and mCherry signal was measured in transiently transfected protoplasts. GFPhigh signal intensity is normalized by mCherry signal intensity and the full length p*lhcsr1* was set to 100%. Average value is marked by a dot (3–4 independent transfection experiments). Asterisks mark significant difference from the full length p*lhcsr1* (*$p < 0.05$, Student's *t*-test). Left: p*lhcsr1* promoter fragments (A–H) with sizes in base pairs. Regions I–IV based on restriction enzyme sites are marked by dashed lines. *lhcsr1* gene promoter (p*lhcsr1*), codon adapted GFP variant (GFPhigh), *35S* terminator (ter), and 2x *35S* promoter (2x p*35S*). The transcription start site predicted by the Plant Promoter Database (Hieno et al., 2014) is marked in the full length p*lhcsr1* promoter with a dashed line.

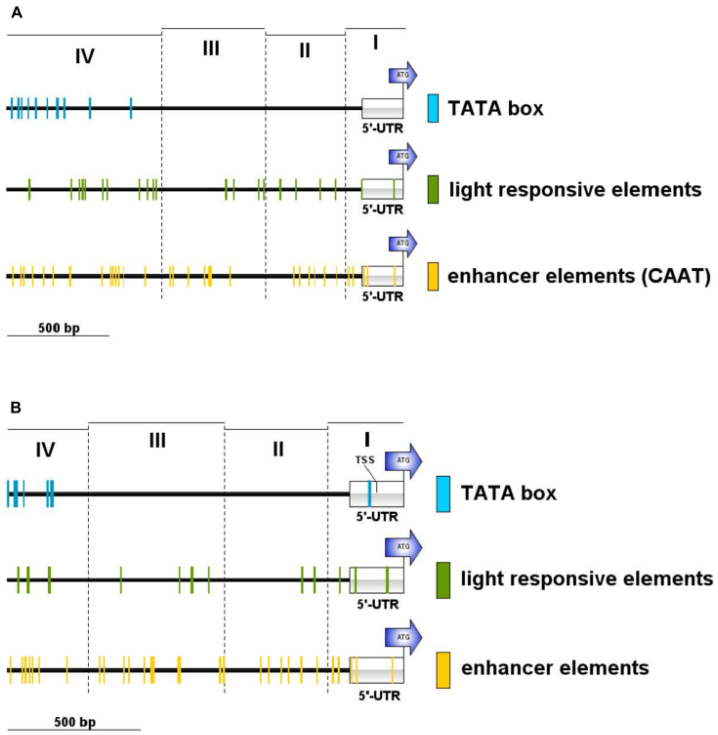



**FIGURE 5 |** Motif occurrences in the *P. patens lhcsr1* promoter. The promoter sequence was screened against the plant *cis*-acting regulatory DNA elements database PLACE **(A)** and the plant *cis*-acting regulatory element database PlantCARE **(B)**. The 1,956 bp upstream region of the *lhcsr1* gene was split into four regions based restriction enzyme sites (*cf.* **Figure 4**). The four regions are marked by dashed lines. Colored lines mark occurrence of light responsive elements (green), TATA box motifs (blue), and enhancer element motifs (yellow). Motifs situated on the plus and minus strand are shown. Position of the 5′ untranslated region (5′-UTR) and the start codon of the *lhcsr1* gene (ATG) are shown. PlantCARE database only used the first 1.5 kbp of the promoter sequence. Regions I and II need to be present to achieve full promoter activity.

(TSS) data implemented in the Plant Promoter database (Hieno et al., 2014) support this TATA box based on 5′ end sequencing of the *lhcsr1* gene with a TSS found 102 bases upstream of the start codon (Supplementary Figure 9). This TATA box is further supported by a possible TSS present in 5′ Cap-capture sequencing data available at the *P. patens* CoGe representation (**Figure 6**).

## DISCUSSION

Promoter:reporter gene constructs can be used to compare expression strength between different promoters. The signal can be read out after cell lysis (Horstmann et al., 2004) or directly from living cells (Thevenin et al., 2012). The readout system established in this study measures the fluorescence signal of transiently transfected protoplasts expressing the reporter genes GFP and mCherry. We use a comparatively low-priced multi-well, multi-wavelength plate reader to measure the fluorescence signal of living protoplasts, while a recent study used a FACS system for measurements to investigate transcription factor activities in living cells (Thevenin et al., 2012). Since transformation efficiencies vary a lot between different transformations and to account for the uptake of varying numbers of plasmids into one cell (Supplementary Figure 2), we introduced a mCherry normalization cassette into all of our constructs. This novel readout system reduces the variation we see in our measurements and thereby helps to differentiate between different promoter strength. This system also allows to operate on lower numbers of protoplasts and biological replicates. We assume that detectable fluorescence correlated in linear fashion with protein amount, an assumption supported by the measurement of protoplasts of stable expressing lines (Supplementary Figure 3).

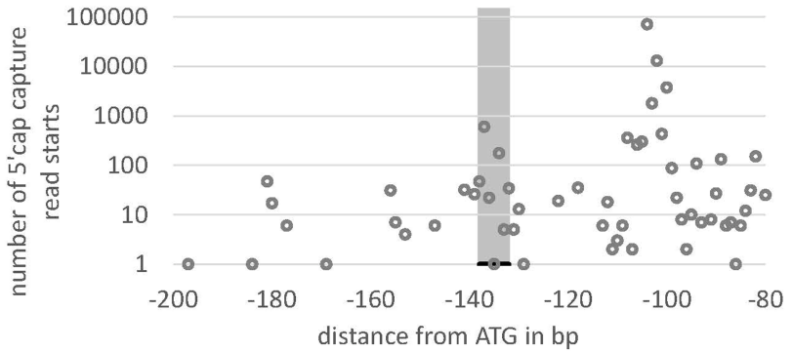



**FIGURE 6 |** Read starting position from 5′ Cap-capture sequencing data for *lhcsr1*. Number of reads (*y*-axis) of RNA-Seq data from *P. patens* gametophores that start within the 5′-UTR at the position –80 to –200 bp upstream of the translation start (ATG) of the *lhcsr1* gene. Gray bar marks TATA box predicted by PlantCARE. Note the peaks at –100 and at the TATA box, potentially representing two transcription start sites (TSS). The upstream TSS at –100 is the one present in the Plant Promoter Database (*cf.* **Figure 4**).

Codon usage of *P. patens* has been calculated and published based on expressed sequence tag (EST) data that were created before the genome sequence was published (Rensing et al., 2005; Stenoien, 2005) and recently based on the gene models predicted from genome sequencing (Szovenyi et al., 2017). In our analysis of codon usage bias based on v1.6 gene models in *P. patens*, we detect a bias in highly expressed genes to prefer certain codons. The same preferences can be seen in other organisms (**Figure 2B**). We can detect a bias for the first codons already within the top 7.3% expressed genes (Supplementary File 1). These highly expressed genes can be confirmed using microarray data (Supplementary Figure 5), and show a bias toward higher GC content and lower ENC (Supplementary Figure 6). In *C. reinhardtii*, codon usage, rather than GC content, was found as key determinant of gene expression efficiency. However, the nucleotide composition was found to feed back to the chromatin state of the promoter region (Barahimipour et al., 2015). To show that the detected codon bias, acting on translation, can be used for biotechnological applications, we created two GFP versions with adapted codon usage but identical amino acid sequence. Starting from the soluble-modified GFP version that gives high signals in particle bombardment transformation assays in *P. patens* (Cho et al., 1999), we exchanged codons to create a GFP that contains codons biased toward high protein expression (GFP high) and a GFP version that contains the opposite codons (GFP low). With the GFP version using preferred codons, we measure higher GFP signals, similar to studies in maize using optimal human codons (Chiu et al., 1996). The effect of our codon changes on protein expression can already be measured when using the 2x *35S* promoter that gives medium expression strength in *P. patens*, and is more pronounced under the control of the stronger *lhcsr1* promoter (**Figure 3**).

Since the 2x *35S* promoter is not the best choice for strong gene expression in biotechnological protein production in *P. patens*, we selected a candidate promoter based on microarray analysis covering important tissues, several stress conditions, and hormone treatments. This includes typical production conditions like protonemal tissue and protoplasts in light. We selected the promoter of *lhcsr1*, a gene coding for a cab protein. LHCSR1 and LHCSR2 are both part of the light harvesting complex and are activated by light. Even though they are mainly expressed during conditions with photosynthetically active tissue, we see a medium RNA expression level under dark conditions in our microarray data (Supplementary Figure 8). This medium expression in darkness should allow the use of this promoter in selection cassettes where diurnal expression is preferred to allow for the upkeep of resistance. In contrast, the single *35S* promoter was found to be inactive in darkness (Saidi et al., 2009). Compared to the 2x *35S* and the *actin1* promoter we see a 2- to 6-fold higher expression when using the *lhcsr1* promoter, based on transcriptional activation. In our expression and readout system, the 2x *35S* and *actin1* promoters do not show significant difference in expression strength, contrary to previous reports in which the *actin1* promoter reached 1.6 times the expression level of the 2x *35S* promoter in a firefly luciferase reporter system (Horstmann et al., 2004). This difference could be due to the different normalization method that uses a second, separate plasmid as transfection control. This resulted in higher variation of the relative luciferase activity between different transfections with a median of coefficients of variation at 0.42 (Supplementary Table 4). Our GFP reporter system with the normalization cassette included into the same plasmid as the promoter of interest provides less variance with the median of coefficients of variation at 0.22 (Supplementary Table 5).

As putative *lhcsr1* promoter sequence, we selected the 1,956 bp upstream of the ATG of the *lhcsr1* gene. Since biotechnological applications will benefit from a shorter promoter sequence, we created shortened versions of the original sequence. We divided the original fragment into four regions based on restriction sites and were able to identify a promoter fragment of 677 bp that

shows full promoter activity. Shorter fragments and fragments that lack the region II (**Figure 4**) show a weaker promoter activity. Removing the 5′-UTR from the 677 bp fragment also results in a low promoter activity, suggesting a regulatory role for the 5′-UTR. Analyzing the promoter fragments for binding sites, PlantCARE detects a TATA box in the 5′-UTR as well as sites known to enhance RNA expression and several binding sites for light regulated transcription factors. With our fully active 677-bp promoter fragment showing at least twofold expression strength of the 2x *35S* promoter, we provide a useful alternative for strong protein expression in *P. patens*. In combination with our codon optimized GFPhigh the 2x *35S* promoter shows a twofold expression strength increase as compared to the 2x *35S* promoter with non-optimized GFP. However, the combination of *lhcsr1* promotor and GFPhigh led to the highest observed activity with up to sixfold expression rate relative to the before mentioned standard.

Our study shows that a significant increase in protein production can be achieved by using suitable combinations of promoter and codon optimization, tackling transcription as well as translation efficiency. In contrast, combining even a strong promoter with genes not codon-optimized for high expression leads to low expression (**Figure 3**). These findings will be especially helpful in biotechnological and proteomics applications producing proteins in the moss *P. patens*, but also to drive, e.g., fluorescence tags and selectable markers. For these future applications, the generation of stable mutant lines with the expression cassettes integrated into the genome will be necessary to test whether the strong expression can also be seen in the genomic context. This will also allow to evaluate the expression strength in different tissues and conditions.

## MATERIALS AND METHODS

### Plant Material

*Physcomitrella patens* Gransden (Rensing et al., 2008) was cultivated on solidified [1% (w/v) agar] mineral medium, also known as modified (Reski and Abel, 1985) Knop's (1868) medium, on 9-cm petri dishes enclosed by laboratory film at 22°C with a 16-h-light/8-h-dark regime under 70 $\mu$mol $m^{-2}$ $s^{-1}$ white light (long-day conditions).

### Codon Usage

To assess codon usage bias, the coding sequences of v1.2 gene models on the Combimatrix microarray experiments from Hiss et al. (2014) that show expression values above the detection limit were analyzed (26,856 genes). Genes were sorted according to their expression level and genes with a normalized fluorescence intensity above 200,000 (567 genes, 2.1%) were termed strongly expressed genes whereas a value of 450,000 (233 genes, 0.87%) put them into the group of highly expressed genes. The codon frequencies for each group were calculated with *R* and afterward a Fisher's Exact Test was used for each of the 64 codons to find significant changes in codon frequencies between the groups. To account for the high number of statistical tests, multiple testing adjusted *p*-values were calculated with the *R* function *p*.adjust and the significance level was set at adjusted $p < 0.05$.

The assessment at which expression level a codon usage bias can be detected was based on the coding sequences of v1.6 gene models with data based on the v1.2 Combimatrix microarray experiments. Custom software used to calculate codon usage and adjusted *p*-values can be found in the Github repository https://github.com/kullrich/bio-scripts/tree/master/codonusage.

### Generation of Transient Constructs

Plasmids for transient transfection were assembled by Golden Gate Cloning using the *Sap*I enzyme. A modified pAM-PAT vector (accession number: AY436765) was used for cloning of 2x p*35S* constructs. The multiple cloning site (MCS) was removed, and together with a chloramphenicol resistance gene and a ccdb kill cassette, two *Sap*I restriction sites were introduced to create the vector pAM-PAT-*Sap*I. To accommodate the insertion of other promoters, the pAM-PAT-*Sap*I vector was modified by restriction with *Xho*I and *Sal*I, removing the 2x *35S* promoter. For cloning reactions promoter and reporter sequences were amplified with primers containing the *Sap*I restriction site with matching overhang. *Sap*I-cut vector, amplified promoter and amplified reporter were incubated with *Sap*I and T4 ligase in ligase buffer and after 1 h incubation at room temperature transformed into TOP10 *E. coli* cells. The mCherry gene was amplified from a modified p123 vector (kindly provided by Michael Bölker) with *Sap*I restriction sites as overhang (mCherry-SapI_fwd, mCherry-SapI_rev) and inserted into the pAM-PAT vector behind the 2x p*35S* sequence. The 2xp*35S*:mCherry fusion was amplified with *Not*I restriction sites as overhang (*Not*I-p35SS_fwd, *Not*I-mCherry) and ligated into the pAM-PAT vector *via* the *Not*I site. Although insertion was not directed by specific restriction sites the clones obtained always contained the 2x p*35S*:mCherry in the same direction as the promoter:GFP insert. To test whether a read-through from the *lhcsr1* promoter could lead to increased expression of mCherry, we removed the 2x *35S* promoter in front of the mCherry. We could not detect a mCherry signal above background when using these constructs (Supplementary Table 6). Plasmid DNA was extracted by the Bibdo protocol (Birnboim and Doly, 1979) or with the NucleoBond Xtra Midi Kit (Macherey-Nagel, Germany).

Modified GFP versions were synthetized by Geneart (Germany) and delivered in the pMA-T vector, inserted *via* the *Sfi*I restriction site. The GFP versions were amplified with primers containing *Sap*I restriction sites as overhang (GFP_high-for_SAP, GFP_high_rev, GFP_low_for, GFP_high_rev). During primer design, the second codon of the smGFP was changed from serine (S) to valine (V). To test whether this has an effect on expression, we compared the GFP/mCherry ratio of the 2x p*35S*:S-V_smGFP to a correct 2x p*35S*:smGFP. The plate reader measurements did not show a difference between the S and V versions of the smGFP (Supplementary Figure 10). 1,956 bp of genomic sequence upstream of the coding sequence for the cab protein (Phypa_169593) were amplified by PCR with primers containing the *Sap*I restriction site as overhang (p169593_for, p169593_rev). Shortened versions of the *lhcsr1* gene promoter sequence were created by restriction digest and blunt/compatible

end ligation with *Eco*RV + *Sal*I (A), *Eco*RV + *Bcl*I (B), *Dra*II (C), StuI (D), *Bcl*I + *Bgl*II (E), *Eco*RV + *Bgl*II (F), and *Sal*I + *Xho*I (G). The *Eco*RV + *Sal*I construct without 5′-UTR (H) was amplified from the full length construct with *Sap*I restriction sites as overhang (sh_pCAB_for, sh_pCAB_rev). The rice *actin1* (McElroy et al., 1990) promoter sequence was amplified by PCR from the PIG-AN vector (Schallenberg-Rudinger et al., 2017) with primers containing the *Sap*I restriction site as overhang (pActin_Sap_for, pActin_Sap_rev).

Cloning success was tested by selection on ampicillin and either by colony PCR or test digestion. Positive candidate plasmids were sent for Sanger sequencing to GATC (Konstanz, Germany) or Macrogen (Amsterdam, Netherlands). The primers used for plasmid construction are shown in Supplementary Table 6.

## Moss Protoplast Transfection

Transfection protocol was adapted from (Hohe et al., 2004). Regularly disrupted protonemal tissue in a 200 mL liquid culture, pH 5.8 was adjusted to 60 mg/L dry weight and transferred to 200 mL liquid Knop medium pH 4.5. After 5–6 days, the culture was harvested by sieving (100 μm sieve). Protonemal tissue was equilibrated in 12 mL 0.51M Mannitol (pH 5.6–5.8) for 30 min, 4 mL Driselase solution (4%) was added and incubated for 1–2 h on a slowly tumbling shaker. The protoplast solution was sieved first on a 100 μm, then on a 50 μm sieve and afterward centrifuged 10 min at 50 *g*. Supernatant was removed, protoplasts resuspended in 10 mL 0.51 M Mannitol and centrifuged 10 min at 50 *g*. Supernatant was removed, protoplasts resuspended in 10 mL 0.51 M Mannitol and protoplast number counted on a Fuchs-Rosenthal counting chamber. Protoplast suspension was centrifuged for 10 min at 50 *g*, supernatant removed and a concentration of $1.2 \times 10^6$ protoplasts per mL adjusted with MMM medium [MMM medium, 0.51 M Mannitol, 15 mM $MgCl_2$, 0.1% w/v 2-(N-morpholino)ethanesulfonic acid, pH 5.6]. For transfection, 100 μL DNA in 0.1 M $Ca(NO_3)_2$, 250 μL protoplast suspension, and 350 μL PEG solution (40% PEG 400 in MMM medium) were mixed during a 30 min incubation time. To slowly dilute the transfection solution, first 1 mL of MMM medium is added and mixed, next 2, 3, 4, and 5 mL are added and mixed every 5 min. Suspension was centrifuged for 10 min at 50 *g* and protoplasts resuspended in regeneration medium (0.28 M glucose and 0.24 M mannitol in Knop medium, pH 5.8). For transient transfections, circular plasmid was used. DNA amounts used for transient transfections were 10–50 μg. After transfection, protoplasts were left to regenerate in 1 mL of regeneration medium.

## Plate Reader Measurements

Fluorescence intensity measurements were performed in a FLUOstar microplate reader (BMG Labtech, Germany) with transfected *P. patens* protoplasts. Sample volumes of 100 μL (up to 30,000 protoplasts) were placed into black 96-well microplates with transparent bottom (Greiner Bio-one, Austria). The samples were detected using the bottom optic, orbital averaging with 2 mm diameter and 15 flashes per well. For GFP and mCherry fluorescence, emission filters at 485 and 584 nm as well as excitation filters at 520 and 620 nm were used with 10 nm bandpass width. Regeneration medium was used as the blank value and non-transfected protoplasts as background control. Blank values were subtracted and the ratio of GFP and mCherry signals was calculated for each well to normalize for the transfection efficiency. Measurements were done 6–7 days after transfection since time course experiments found the highest fluorescence signal intensity after this time period (Supplementary Figure 11).

## Gene Ontology (GO) Analyses and Visualization

The GO bias analyses used Fisher's Exact Test to calculate *p*-values. Multiple testing corrected (Benjamini and Hochberg, 1995) *q*-values were calculated in *R* with the function *p*.adjust (R Development Core Team, 2008). Word cloud visualizations were created using the online tool wordle[1] . Word size is proportional to the −log10(*q*-value) and over-represented GO terms were colored dark green if $q \leq 0.0001$ and light green if $q > 0.0001$. Under-represented GO terms were colored dark red if $q \leq 0.0001$ and light red if $q > 0.0001$.

## AUTHOR CONTRIBUTIONS

CG, CN, LS, MB, and MH prepared the constructs, transfected protoplasts, and analyzed data. AS and KU calculated codon frequencies. SR conceived of the work. MS-R, P-FP, and SR supervised the project. MH, MS-R, P-FP, and SR designed the experiments. MH and SR wrote the manuscript with contributions by all authors.

## FUNDING

This work was supported by the German Federal Ministry of Education and Research (Freiburg Initiative for Systems Biology, 0313921 to SR).

## ACKNOWLEDGMENTS

We want to thank the students Christian Peikert, Christian Volk, Stefan Ost, and Anne Genau who helped with cloning, transfection, and bioinformatics analysis. We want to thank Stefanie Pilz, Marco Göttig, and Faezeh Donges for their excellent technical assistance. We thank Michael Bölker for providing the p123 vector.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2017.01842/full#supplementary-material

[1] http://www.wordle.net/

## REFERENCES

Abe, J., Hiwatashi, Y., Ito, M., Hasebe, M., and Sekimoto, H. (2008). Expression of exogenous genes under the control of endogenous HSP70 and CAB promoters in the *Closterium peracerosum*-strigosum-littorale complex. *Plant Cell Physiol.* 49, 625–632.

Alboresi, A., Gerotto, C., Giacometti, G. M., Bassi, R., and Morosinotto, T. (2010). *Physcomitrella patens* mutants affected on heat dissipation clarify the evolution of photoprotection mechanisms upon land colonization. *Proc. Natl. Acad. Sci. U.S.A.* 15, 11128–11133. doi: 10.1073/pnas.1002873107

Barahimipour, R., Strenkert, D., Neupert, J., Schroda, M., Merchant, S. S., and Bock, R. (2015). Dissecting the contributions of GC content and codon usage to gene expression in the model alga *Chlamydomonas reinhardtii*. *Plant J.* 84, 704–717. doi: 10.1111/tpj.13033

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.* 57, 289–300.

Bezanilla, M., Pan, A., and Quatrano, R. S. (2003). RNA interference in the moss *Physcomitrella patens*. *Plant Physiol.* 133, 470–474. doi: 10.1104/pp.103.024901

Birnboim, H. C., and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* 24, 1513–1523. doi: 10.1093/nar/7.6.1513

Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V., and Kondrashov, F. A. (2002). Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418. doi: 10.1038/ng940

Chiu, W., Niwa, Y., Zeng, W., Hirano, T., Kobayashi, H., and Sheen, J. (1996). Engineered GFP as a vital reporter in plants. *Curr. Biol.* 6, 325–330. doi: 10.1016/S0960-9822(02)00483-9

Cho, S. H., Chung, Y. S., Cho, S. K., Rim, Y. W., and Shin, J. S. (1999). Particle bombardment mediated transformation and GFP expression in the moss *Physcomitrella patens*. *Mol. Cells* 28, 14–19.

Christensen, A. H., Sharrock, R. A., and Quail, P. H. (1992). Maize polyubiquitin genes: structure, thermal perturbation of expression and transcript splicing, and promoter activity following transfer to protoplasts by electroporation. *Plant Mol. Biol.* 18, 675–689. doi: 10.1007/BF00020010

Cormack, B. P., Bertram, G., Egerton, M., Gow, N. A., Falkow, S., and Brown, A. J. (1997). Yeast-enhanced green fluorescent protein (yEGFP): a reporter of gene expression in *Candida albicans*. *Microbiology* 143(Pt 2), 303–311. doi: 10.1099/00221287-143-2-303

Davis, S. J., and Vierstra, R. D. (1998). Soluble, highly fluorescent variants of green fluorescent protein (GFP) for use in higher plants. *Plant Mol. Biol.* 36, 521–528. doi: 10.1023/A:1005991617182

Dittami, S. M., Michel, G., Collen, J., Boyen, C., and Tonon, T. (2010). Chlorophyll-binding proteins revisited–a multigenic family of light-harvesting and stress proteins from a brown algal perspective. *BMC Evol. Biol.* 10:365. doi: 10.1186/1471-2148-10-365

dos Reis, M., and Wernisch, L. (2009). Estimating translational selection in eukaryotic genomes. *Mol. Biol. Evol.* 26, 451–461. doi: 10.1093/molbev/msn272

Duret, L., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 13, 4482–4487. doi: 10.1073/pnas.96.8.4482

Fuhrmann, M., Oertel, W., and Hegemann, P. (1999). A synthetic gene coding for the green fluorescent protein (GFP) is a versatile reporter in *Chlamydomonas reinhardtii*. *Plant J.* 19, 353–361. doi: 10.1046/j.1365-313X.1999.00526.x

Gerotto, C., Alboresi, A., Giacometti, G. M., Bassi, R., and Morosinotto, T. (2011). Role of PSBS and LHCSR in *Physcomitrella patens* acclimation to high light and low temperature. *Plant Cell Environ.* 34, 922–932. doi: 10.1111/j.1365-3040.2011.02294.x

Giuliano, G., Pichersky, E., Malik, V. S., Timko, M. P., Scolnik, P. A., and Cashmore, A. R. (1988). An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene. *Proc. Natl. Acad. Sci. U.S.A.* 85, 7089–7093. doi: 10.1073/pnas.85.19.7089

Green, P. J., Yong, M. H., Cuozzo, M., Kano-Murakami, Y., Silverstein, P., and Chua, N. H. (1988). Binding site requirements for pea nuclear protein factor GT-1 correlate with sequences required for light-dependent transcriptional activation of the rbcS-3A gene. *EMBO J.* 20, 4035–4044.

Hieno, A., Naznin, H. A., Hyakumachi, M., Sakurai, T., Tokizawa, M., Koyama, H., et al. (2014). ppdb: plant promoter database version 3.0. *Nucleic Acids Res.* 42, D1188–D1192. doi: 10.1093/nar/gkt1027

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* 27, 297–300. doi: 10.1093/nar/27.1.297

Hiraoka, Y., Kawamata, K., Haraguchi, T., and Chikashige, Y. (2009). Codon usage bias is correlated with gene expression levels in the fission yeast *Schizosaccharomyces pombe*. *Genes Cells* 14, 499–509. doi: 10.1111/j.1365-2443.2009.01284.x

Hiss, M., Laule, O., Meskauskiene, R. M., Arif, M. A., Decker, E. L., Erxleben, A., et al. (2014). Large-scale gene expression profiling data for the model moss *Physcomitrella patens* aid understanding of developmental progression, culture and stress conditions. *Plant J.* 79, 530–539. doi: 10.1111/tpj.12572

Hohe, A., Egener, T., Lucht, J. M., Holtorf, H., Reinhard, C., Schween, G., et al. (2004). An improved and highly standardised transformation procedure allows efficient production of single and multiple targeted gene-knockouts in a moss, *Physcomitrella patens*. *Curr. Genet.* 44, 339–347. doi: 10.1007/s00294-003-0458-4

Horstmann, V., Huether, C. M., Jost, W., Reski, R., and Decker, E. L. (2004). Quantitative promoter analysis in *Physcomitrella patens*: a set of plant vectors activating gene expression within three orders of magnitude. *BMC Biotechnol.* 4:13. doi: 10.1186/1472-6750-4-13

Jost, W., Link, S., Horstmann, V., Decker, E. L., Reski, R., and Gorr, G. (2005). Isolation and characterisation of three moss-derived beta-tubulin promoters suitable for recombinant expression. *Curr. Genet.* 47, 111–120. doi: 10.1007/s00294-004-0555-z

Knop, W. (1868). *Der Kreislauf des Stoffs: Lehrbuch der Agricultur-Chemie*. Leipzig: H. Haessel.

Komar, A. A. (2016). The Yin and Yang of codon usage. *Hum. Mol. Genet.* 25, R77–R85.

Kondo, T., Pinnola, A., Chen, W. J., Dall'Osto, L., Bassi, R., and Schlau-Cohen, G. S. (2017). Single-molecule spectroscopy of LHCSR1 protein dynamics identifies two distinct states responsible for multi-timescale photosynthetic photoprotection. *Nat. Chem.* 9, 772–778. doi: 10.1038/nchem.2818

Koziol, A. G., Borza, T., Ishida, K., Keeling, P., Lee, R. W., and Durnford, D. G. (2007). Tracing the evolution of the light-harvesting antennae in chlorophyll a/b-containing organisms. *Plant Physiol.* 143, 1802–1816. doi: 10.1104/pp.106.092536

Kubo, M., Imai, A., Nishiyama, T., Ishikawa, M., Sato, Y., Kurata, T., et al. (2013). System for stable beta-estradiol-inducible gene expression in the moss *Physcomitrella patens*. *PLOS ONE* 8:e77356. doi: 10.1371/journal.pone.0077356

Li, X. P., Bjorkman, O., Shih, C., Grossman, A. R., Rosenquist, M., Jansson, S., et al. (2000). A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature* 27, 391–395. doi: 10.1038/35000131

Liu, W., Xie, Y., Ma, J., Luo, X., Nie, P., Zuo, Z., et al. (2015). IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* 15, 3359–3361. doi: 10.1093/bioinformatics/btv362

Luan, S., and Bogorad, L. (1992). A rice cab gene promoter contains separate cis-acting elements that regulate expression in dicot and monocot plants. *Plant Cell* 4, 971–981. doi: 10.1105/tpc.4.8.971

McElroy, D., Zhang, W., Cao, J., and Wu, R. (1990). Isolation of an efficient actin promoter for use in rice transformation. *Plant Cell* 2, 163–171. doi: 10.1105/tpc.2.2.163

Morton, B. R., and Wright, S. I. (2007). Selective constraints on codon usage of nuclear genes from *Arabidopsis thaliana*. *Mol. Biol. Evol.* 24, 122–129. doi: 10.1093/molbev/msl139

Ortiz-Ramirez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., et al. (2015). A transcriptome atlas of *Physcomitrella patens* provides insights into the evolution and development of land plants. *Mol. Plant* 11, 205–220. doi: 10.1016/j.molp.2015.12.002

Perroud, P. F., Cove, D. J., Quatrano, R. S., and McDaniel, S. F. (2011). An experimental method to facilitate the identification of hybrid sporophytes in the moss *Physcomitrella patens* using fluorescent tagged lines. *New Phytol.* 191, 301–306. doi: 10.1111/j.1469-8137.2011.03668.x

Pinnola, A., Ballottari, M., Bargigia, I., Alcocer, M., D'Andrea, C., Cerullo, G., et al. (2017). Functional modulation of LHCSR1 protein from *Physcomitrella patens*

by zeaxanthin binding and low pH. *Sci. Rep.* 11, 11158. doi: 10.1038/s41598-017-11101-7

Quax, T. E., Claassens, N. J., Soll, D., and van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. *Mol. Cell* 16, 149–161. doi: 10.1016/j.molcel.2015.05.035

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna: The R Foundation for Statistical Computing.

Rensing, S. A., Fritzowsky, D., Lang, D., and Reski, R. (2005). Protein encoding genes in an ancient plant: analysis of codon usage, retained genes and splice sites in a moss, *Physcomitrella patens*. *BMC Genomics* 6:43.

Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., et al. (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* 319, 64–69. doi: 10.1126/science.1150646

Reski, R., and Abel, W. O. (1985). Induction of budding on chloronemata and caulonemata of the moss, physcomitrella-patens, using isopentenyladenine. *Planta* 165, 354–358. doi: 10.1007/BF00392232

Saidi, Y., Finka, A., Chakhporanian, M., Zryd, J. P., Schaefer, D. G., and Goloubinoff, P. (2005). Controlled expression of recombinant proteins in *Physcomitrella patens* by a conditional heat-shock promoter: a tool for plant research and biotechnology. *Plant Mol. Biol.* 59, 697–711. doi: 10.1007/s11103-005-0889-z

Saidi, Y., Schaefer, D. G., Goloubinoff, P., Zryd, J. P., and Finka, A. (2009). The CaMV 35S promoter has a weak expression activity in dark grown tissues of moss *Physcomitrella patens*. *Plant Signal. Behav.* 4, 457–459. doi: 10.4161/psb.4.5.8541

Schaefer, D., Zryd, J. P., Knight, C. D., and Cove, D. J. (1991). Stable transformation of the moss *Physcomitrella patens*. *Mol. Gen. Genet.* 226, 418–424. doi: 10.1007/BF00260654

Schallenberg-Rudinger, M., Oldenkott, B., Hiss, M., Trinh, P. L., Knoop, V., and Rensing, S. A. (2017). A single-target mitochondrial RNA editing factor of *Funaria hygrometrica* can fully reconstitute RNA Editing at two sites in *Physcomitrella patens*. *Plant Cell Physiol.* 01, 496–507. doi: 10.1093/pcp/pcw229

Stenoien, H. K. (2005). Adaptive basis of codon usage in the haploid moss *Physcomitrella patens*. *Heredity* 94, 87–93. doi: 10.1038/sj.hdy.6800547

Szovenyi, P., Ullrich, K. K., Rensing, S. A., Lang, D., van Gessel, N., Stenoien, H. K., et al. (2017). Selfing in haploid plants and efficacy of selection: codon usage bias in the model moss *Physcomitrella patens*. *Genome Biol. Evol.* 9, 1528–1546.

Thevenin, J., Dubos, C., Xu, W., Le Gourrierec, J., Kelemen, Z., Charlot, F., et al. (2012). A new system for fast and quantitative analysis of heterologous gene expression in plants. *New Phytol.* 193, 504–512. doi: 10.1111/j.1469-8137.2011.03936.x

Ullrich, K. K., Hiss, M., and Rensing, S. A. (2015). Means to optimize protein expression in transgenic plants. *Curr. Opin. Biotechnol.* 32, 61–67. doi: 10.1016/j.copbio.2014.11.011

van Hemert, F. J., and Berkhout, B. (1995). The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. *J. Mol. Evol.* 41, 132–140. doi: 10.1007/BF00170664

Weise, A., Rodriguez-Franco, M., Timm, B., Hermann, M., Link, S., Jost, W., et al. (2006). Use of *Physcomitrella patens* actin 5' regions for high transgene expression: importance of 5' introns. *Appl. Microbiol. Biotechnol.* 70, 337–345. doi: 10.1007/s00253-005-0087-6

Widiez, T., Symeonidi, A., Luo, C., Lam, E., Lawton, M., and Rensing, S. A. (2014). The chromatin landscape of the moss *Physcomitrella patens* and its dynamics during development and drought stress. *Plant J.* 79, 67–81. doi: 10.1111/tpj.12542

Wolf, L., Rizzini, L., Stracke, R., Ulm, R., and Rensing, S. A. (2010). The molecular and physiological responses of *Physcomitrella patens* to ultraviolet-B radiation. *Plant Physiol.* 153, 1123–1134. doi: 10.1104/pp.110.154658

Wright, S. I., Yau, C. B., Looseley, M., and Meyers, B. C. (2004). Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Mol. Biol. Evol.* 21, 1719–1726. doi: 10.1093/molbev/msh191

Zemach, A., McDaniel, I. E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 14, 916–919. doi: 10.1126/science.1186366

Zimmer, A. D., Lang, D., Buchta, K., Rombauts, S., Nishiyama, T., Hasebe, M., et al. (2013). Reannotation and extended community resources for the genome of the non-seed plant *Physcomitrella patens* provide insights into the evolution of plant gene structures and functions. *BMC Genomics* 14:498. doi: 10.1186/1471-2164-14-498

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# 7 Concluding remarks

In my thesis I could show that the *Physcomitrella patens* ecotype from Reute is a versatile tool for molecular biology and shows a high fertility. It is therefore especially suited for studies concerned with sporophyte development and is genetically closer to the commonly used ecotype Gransden than the Villersexel K3 ecotype. I further could show that transient transformation is possible in the Reute ecotype. The experiments indicate that expression data from one ecotype can be used as proxy for the expression strength in another ecotype by calculating the expression strength based on expression data covering a broad range of tissues and treatments. Further my thesis contains data on applied codon-usage optimization in *Physcomitrella patens* and shows the feasibility of gene expression optimization by using a suitable promoter and codon usage combination. My work resulted in the creation of a codon-optimized GFP version that is available to the moss community.

In summary the findings from my publications characterize the *P. patens* ecotype Reute and demonstrate that it will be a useful tool in reverse genetic studies.

# 8 Cited references

**Aharoni, A. and Vorst, O.** (2002) DNA microarrays for functional plant genomics. *Plant Mol Biol*, **48**, 99-118.

**Arabidopsis Genome, I.** (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796-815.

**Beike, A.K., Horst, N.A. and Rensing, S.A.** (2010) Axenic bryophyte in vitro cultivation. *Journal of Endocytobiosis and Cell Research*, **20**, 7.

**Beike, A.K., Lang, D., Zimmer, A.D., Wust, F., Trautmann, D., Wiedemann, G., Beyer, P., Decker, E.L. and Reski, R.** (2015) Insights from the cold transcriptome of Physcomitrella patens: global specialization pattern of conserved transcriptional regulators and identification of orphan genes involved in cold acclimation. *New Phytol*, **205**, 869-881.

**Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara, E.C.M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R. and Smith, A.J.** (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53-59.

**Blanchard, A.P. and Hood, L.** (1996) Sequence to array: probing the genome's secrets. *Nat Biotechnol*, **14**, 1649.

**Busch, H., Boerries, M., Bao, J., Hanke, S.T., Hiss, M., Tiko, T. and Rensing, S.A.** (2013) Network theory inspired analysis of time-resolved expression data reveals key players guiding P. patens stem cell development. *PLoS One*, **8**, e60494.

**Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Muller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K.J. and Weigel, D.** (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet*, **43**, 956-963.

**Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. and Prasher, D.C.** (1994) Green fluorescent protein as a marker for gene expression. *Science*, **263**, 802-805.

**Chen, Y.R., Su, Y.S. and Tu, S.L.** (2012) Distinct phytochrome actions in nonvascular plants revealed by targeted inactivation of phytobilin biosynthesis. *P Natl Acad Sci USA*, **109**, 8310-8315.

**Cho, S.H., Chung, Y.S., Cho, S.K., Rim, Y.W. and Shin, J.S.** (1999) Particle bombardment mediated transformation and GFP expression in the moss Physcomitrella patens. *Mol Cells*, **9**, 14-19.

**Cove, D.J., Perroud, P.F., Charron, A.J., McDaniel, S.F., Khandelwal, A. and Quatrano, R.S.** (2009) Isolation and regeneration of protoplasts of the moss Physcomitrella patens. *Cold Spring Harb Protoc*, **2009**, pdb prot5140.

**Crick, F.H., Barnett, L., Brenner, S. and Watts-Tobin, R.J.** (1961) General nature of the genetic code for proteins. *Nature*, **192**, 1227-1232.

**Cuming, A.C., Cho, S.H., Kamisugi, Y., Graham, H. and Quatrano, R.S.** (2007) Microarray analysis of transcriptional responses to abscisic acid and osmotic, salt, and drought stress in the moss, Physcomitrella patens. *New Phytol*, **176**, 275-287.

**Davis, S.J. and Vierstra, R.D.** (1998) Soluble, highly fluorescent variants of green fluorescent protein (GFP) for use in higher plants. *Plant Mol Biol*, **36**, 521-528.

**Demko, V., Perroud, P.F., Johansen, W., Delwiche, C.F., Cooper, E.D., Remme, P., Ako, A.E., Kugler, K.G., Mayer, K.F., Quatrano, R. and Olsen, O.A.** (2014) Genetic analysis of DEFECTIVE KERNEL1 loop function in three-dimensional body patterning in Physcomitrella patens. *Plant Physiol*, **166**, 903-919.

**Duret, L. and Mouchiroud, D.** (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *P Natl Acad Sci USA*, **96**, 4482-4487.

**Egener, T., Granado, J., Guitton, M.C., Hohe, A., Holtorf, H., Lucht, J.M., Rensing, S.A., Schlink, K., Schulte, J., Schween, G., Zimmermann, S., Duwenig, E., Rak, B. and Reski, R.** (2002) High frequency of phenotypic deviations in Physcomitrella patens plants transformed with a gene-disruption library. *BMC Plant Biol*, **2**, 6.

**Engel, P.P.** (1968) The Induction of Biochemical and Morphological Mutants in the Moss Physcomitrella patens. *Am J Bot*, **55**, 438-446.

**Engelen, K., Naudts, B., De Moor, B. and Marchal, K.** (2006) A calibration method for estimating absolute expression levels from microarray data. *Bioinformatics*, **22**, 1251-1258.

**Feng, Z., Zhang, B., Ding, W., Liu, X., Yang, D.L., Wei, P., Cao, F., Zhu, S., Zhang, F., Mao, Y. and Zhu, J.K.** (2013) Efficient genome editing in plants using a CRISPR/Cas system. *Cell Res*, **23**, 1229-1232.

**Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T. and Solas, D.** (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767-773.

**Frank, M.H. and Scanlon, M.J.** (2015) Cell-specific transcriptomic analyses of three-dimensional shoot development in the moss Physcomitrella patens. *The Plant journal : for cell and molecular biology*, **83**, 743-751.

**Galbraith, D.W.** (2006) DNA microarray analyses in higher plants. *OMICS*, **10**, 455-473.

**Grantham, R.** (1978) Viral, prokaryote and eukaryote genes contrasted by mRNA sequence indexes. *FEBS Lett*, **95**, 1-11.

**Harrison, C.J., Roeder, A.H., Meyerowitz, E.M. and Langdale, J.A.** (2009) Local cues and asymmetric cell divisions underpin body plan transitions in the moss Physcomitrella patens. *Curr Biol*, **19**, 461-471.

**Heather, J.M. and Chain, B.** (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics*, **107**, 1-8.

**Heim, R., Cubitt, A.B. and Tsien, R.Y.** (1995) Improved green fluorescence. *Nature*, **373**, 663-664.

**Higgs, P.G. and Ran, W.** (2008) Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol Biol Evol*, **25**, 2279-2291.

**Hinnen, A., Hicks, J.B. and Fink, G.R.** (1978) Transformation of yeast. *Proc Natl Acad Sci U S A*, **75**, 1929-1933.

**Hiss, M., Laule, O., Meskauskiene, R.M., Arif, M.A., Decker, E.L., Erxleben, A., Frank, W., Hanke, S.T., Lang, D., Martin, A., Neu, C., Reski, R., Richardt, S., Schallenberg-Rudinger, M.,**

**Szovenyi, P., Tiko, T., Wiedemann, G., Wolf, L., Zimmermann, P. and Rensing, S.A.** (2014) Large-scale gene expression profiling data for the model moss Physcomitrella patens aid understanding of developmental progression, culture and stress conditions. *The Plant journal : for cell and molecular biology*, **79**, 530-539.

**Hohe, A., Rensing, S.A., Mildner, M., Lang, D. and Reski, R.** (2002) Day length and temperature strongly influence sexual reproduction and expression of a novel MADS-Box gene in the moss Physcomitrella patens (vol 4, pg 595, 2002). *Plant Biology*, **4**, 762-762.

**Inouye, S. and Tsuji, F.I.** (1994) Aequorea green fluorescent protein. Expression of the gene and fluorescence characteristics of the recombinant protein. *FEBS Lett*, **341**, 277-280.

**Itakura, K., Hirose, T., Crea, R., Riggs, A.D., Heyneker, H.L., Bolivar, F. and Boyer, H.W.** (1977) Expression in Escherichia coli of a chemically synthesized gene for the hormone somatostatin. *Science*, **198**, 1056-1063.

**Jill Harrison, C.** (2017) Development and genetics in the evolution of land plant body plans. *Philos Trans R Soc Lond B Biol Sci*, **372**.

**Kafatos, F.C., Jones, C.W. and Efstratiadis, A.** (1979) Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic Acids Res*, **7**, 1541-1552.

**Kamisugi, Y., von Stackelberg, M., Lang, D., Care, M., Reski, R., Rensing, S.A. and Cuming, A.C.** (2008) A sequence-anchored genetic linkage map for the moss, Physcomitrella patens. *The Plant journal : for cell and molecular biology*, **56**, 855-866.

**Khraiwesh, B., Qudeimat, E., Thimma, M., Chaiboonchoe, A., Jijakli, K., Alzahmi, A., Arnoux, M. and Salehi-Ashtiani, K.** (2015) Genome-wide expression analysis offers new insights into the origin and evolution of Physcomitrella patens stress response. *Sci Rep*, **5**, 17434.

**Komatsu, K., Suzuki, N., Kuwamura, M., Nishikawa, Y., Nakatani, M., Ohtawa, H., Takezawa, D., Seki, M., Tanaka, M., Taji, T., Hayashi, T. and Sakata, Y.** (2013) Group A PP2Cs evolved in land plants as key regulators of intrinsic desiccation tolerance. *Nat Commun*, **4**, 2219.

**Lang, D., Ullrich, K.K., Murat, F., Fuchs, J., Jenkins, J., Haas, F.B., Piednoel, M., Gundlach, H., Van Bel, M., Meyberg, R., Vives, C., Morata, J., Symeonidi, A., Hiss, M., Muchero, W., Kamisugi, Y., Saleh, O., Blanc, G., Decker, E.L., van Gessel, N., Grimwood, J., Hayes, R.D., Graham, S.W., Gunter, L.E., McDaniel, S.F., Hoernstein, S.N.W., Larsson, A., Li, F.W., Perroud, P.F., Phillips, J., Ranjan, P., Rokshar, D.S., Rothfels, C.J., Schneider, L., Shu, S., Stevenson, D.W., Thummler, F., Tillich, M., Villarreal Aguilar, J.C., Widiez, T., Wong, G.K., Wymore, A., Zhang, Y., Zimmer, A.D., Quatrano, R.S., Mayer, K.F.X., Goodstein, D., Casacuberta, J.M., Vandepoele, K., Reski, R., Cuming, A.C., Tuskan, G.A., Maumus, F., Salse, J., Schmutz, J. and Rensing, S.A.** (2018) The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J*, **93**, 515-533.

**Li, J.F., Norville, J.E., Aach, J., McCormack, M., Zhang, D., Bush, J., Church, G.M. and Sheen, J.** (2013) Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. *Nat Biotechnol*, **31**, 688-691.

**Liu, Y.C. and Vidali, L.** (2011) Efficient polyethylene glycol (PEG) mediated transformation of the moss Physcomitrella patens. *J Vis Exp*.

**Lopez-Obando, M., Hoffmann, B., Gery, C., Guyon-Debast, A., Teoule, E., Rameau, C., Bonhomme, S. and Nogue, F.** (2016) Simple and Efficient Targeting of Multiple Genes Through CRISPR-Cas9 in Physcomitrella patens. *G3 (Bethesda)*, **6**, 3647-3653.

**Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Marechal-Drouard, L., Marshall, W.F., Qu, L.H., Nelson, D.R., Sanderfoot, A.A., Spalding, M.H., Kapitonov, V.V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S.M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C.L., Cognat, V., Croft, M.T., Dent, R., Dutcher, S., Fernandez, E., Fukuzawa, H., Gonzalez-Ballester, D., Gonzalez-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P.A., Lemaire, S.D., Lobanov, A.V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J.V., Moseley, J., Napoli, C., Nedelcu, A.M., Niyogi, K., Novoselov, S.V., Paulsen, I.T., Pazour, G., Purton, S.,**

**Ral, J.P., Riano-Pachon, D.M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S.L., Allmer, J., Balk, J., Bisova, K., Chen, C.J., Elias, M., Gendler, K., Hauser, C., Lamb, M.R., Ledford, H., Long, J.C., Minagawa, J., Page, M.D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A.M., Yang, P., Ball, S., Bowler, C., Dieckmann, C.L., Gladyshev, V.N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R.T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L., Huang, Y.W., Jhaveri, J., Luo, Y., Martinez, D., Ngau, W.C., Otillar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., Grigoriev, I.V., Rokhsar, D.S. and Grossman, A.R.** (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245-250.

**Metzker, M.L.** (2010) Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31-46.

**Nekrasov, V., Staskawicz, B., Weigel, D., Jones, J.D. and Kamoun, S.** (2013) Targeted mutagenesis in the model plant Nicotiana benthamiana using Cas9 RNA-guided endonuclease. *Nat Biotechnol*, **31**, 691-693.

**Nishiyama, T., Miyawaki, K., Ohshima, M., Thompson, K., Nagashima, A., Hasebe, M. and Kurata, T.** (2012) Digital gene expression profiling by 5'-end sequencing of cDNAs during reprogramming in the moss Physcomitrella patens. *PLoS One*, **7**, e36471.

**O'Donoghue, M.T., Chater, C., Wallace, S., Gray, J.E., Beerling, D.J. and Fleming, A.J.** (2013) Genome-wide transcriptomic analysis of the sporophyte of the moss Physcomitrella patens. *J Exp Bot*, **64**, 3567-3581.

**Ortiz-Ramirez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., Feijo, J.A. and Becker, J.D.** (2016) A Transcriptome Atlas of Physcomitrella patens Provides Insights into the Evolution and Development of Land Plants. *Mol Plant*, **9**, 205-220.

**Perroud, P.F., Cove, D.J., Quatrano, R.S. and McDaniel, S.F.** (2011) An experimental method to facilitate the identification of hybrid sporophytes in the moss Physcomitrella patens using fluorescent tagged lines. *New Phytol*, **191**, 301-306.

**Perroud, P.F., Haas, F.B., Hiss, M., Ullrich, K.K., Alboresi, A., Amirebrahimi, M., Barry, K., Bassi, R., Bonhomme, S., Chen, H., Coates, J.C., Fujita, T., Guyon-Debast, A., Lang, D., Lin, J., Lipzen, A., Nogue, F., Oliver, M.J., Ponce de Leon, I., Quatrano, R.S., Rameau, C., Reiss, B., Reski, R., Ricca, M., Saidi, Y., Sun, N., Szovenyi, P., Sreedasyam, A., Grimwood, J., Stacey, G., Schmutz, J. and Rensing, S.A.** (2018) The Physcomitrella patens gene atlas project: large-scale RNA-seq based expression data. *Plant J*, **95**, 168-182.

**Rensing, S.A.** (2017) Why we need more non-seed plant models. *New Phytol*, **216**, 355-360.

**Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin, I.T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S., Yamaguchi, K., Sugano, S., Kohara, Y., Fujiyama, A., Anterola, A., Aoki, S., Ashton, N., Barbazuk, W.B., Barker, E., Bennetzen, J.L., Blankenship, R., Cho, S.H., Dutcher, S.K., Estelle, M., Fawcett, J.A., Gundlach, H., Hanada, K., Heyl, A., Hicks, K.A., Hughes, J., Lohr, M., Mayer, K., Melkozernov, A., Murata, T., Nelson, D.R., Pils, B., Prigge, M., Reiss, B., Renner, T., Rombauts, S., Rushton, P.J., Sanderfoot, A., Schween, G., Shiu, S.H., Stueber, K., Theodoulou, F.L., Tu, H., Van de Peer, Y., Verrier, P.J., Waters, E., Wood, A., Yang, L., Cove, D., Cuming, A.C., Hasebe, M., Lucas, S., Mishler, B.D., Reski, R., Grigoriev, I.V., Quatrano, R.S. and Boore, J.L.** (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64-69.

**Saiki, R.K., Walsh, P.S., Levenson, C.H. and Erlich, H.A.** (1989) Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc Natl Acad Sci U S A*, **86**, 6230-6234.

**Sakakibara, K., Reisewitz, P., Aoyama, T., Friedrich, T., Ando, S., Sato, Y., Tamada, Y., Nishiyama, T., Hiwatashi, Y., Kurata, T., Ishikawa, M., Deguchi, H., Rensing, S.A., Werr, W., Murata, T., Hasebe, M. and Laux, T.** (2014) WOX13-like genes are required for reprogramming of leaf and protoplast cells into stem cells in the moss Physcomitrella patens. *Development*, **141**, 1660-1670.

**Sanchez-Vera, V., Kenchappa, C.S., Landberg, K., Bressendorff, S., Schwarzbach, S., Martin, T., Mundy, J., Petersen, M., Thelander, M. and Sundberg, E.** (2017) Autophagy is required for gamete differentiation in the moss Physcomitrella patens. *Autophagy*, **13**, 1939-1951.

**Schaefer, D., Zryd, J.P., Knight, C.D. and Cove, D.J.** (1991) Stable transformation of the moss Physcomitrella patens. *Mol Gen Genet*, **226**, 418-424.

**Schaefer, D.G. and Zryd, J.P.** (1997) Efficient gene targeting in the moss Physcomitrella patens. *Plant J*, **11**, 1195-1206.

**Schena, M., Shalon, D., Davis, R.W. and Brown, P.O.** (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-470.

**Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J.J., Qiu, J.L. and Gao, C.** (2013) Targeted genome modification of crop plants using a CRISPR-Cas system. *Nat Biotechnol*, **31**, 686-688.

**Sharp, P.M., Stenico, M., Peden, J.F. and Lloyd, A.T.** (1993) Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans*, **21**, 835-841.

**Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A. and Waterston, R.H.** (2017) DNA sequencing at 40: past, present and future. *Nature*, **550**, 345-353.

**Stenoien, H.K.** (2005) Adaptive basis of codon usage in the haploid moss Physcomitrella patens. *Heredity*, **94**, 87-93.

**Strotbek, C., Krinninger, S. and Frank, W.** (2013) The moss Physcomitrella patens: methods and tools from cultivation to targeted analysis of gene function. *Int J Dev Biol*, **57**, 553-564.

**Szovenyi, P., Ullrich, K.K., Rensing, S.A., Lang, D., van Gessel, N., Stenoien, H.K., Conti, E. and Reski, R.** (2017) Selfing in haploid plants and efficacy of selection: codon usage bias in the model moss Physcomitrella patens. *Genome Biol Evol*.

**Tsien, R.Y.** (2005) Building and breeding molecules to spy on cells and tumors. *FEBS Lett*, **579**, 927-932.

**von Stackelberg, M., Rensing, S.A. and Reski, R.** (2006) Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. *BMC plant biology*, **6**, 9.

**Wolf, L., Rizzini, L., Stracke, R., Ulm, R. and Rensing, S.A.** (2010) The molecular and physiological responses of Physcomitrella patens to ultraviolet-B radiation. *Plant Physiol*, **153**, 1123-1134.

**Wu, H.P., Su, Y.S., Chen, H.C., Chen, Y.R., Wu, C.C., Lin, W.D. and Tu, S.L.** (2014) Genome-wide analysis of light-regulated alternative splicing mediated by photoreceptors in Physcomitrella patens. *Genome biology*, **15**, R10.

**Xiao, L., Wang, H., Wan, P., Kuang, T. and He, Y.** (2011) Genome-wide transcriptome analysis of gametophyte development in Physcomitrella patens. *BMC plant biology*, **11**, 177.

**Xiao, L., Zhang, L., Yang, G., Zhu, H. and He, Y.** (2012) Transcriptome of protoplasts reprogrammed into stem cells in Physcomitrella patens. *PLoS One*, **7**, e35961.

**Xie, K. and Yang, Y.** (2013) RNA-guided genome editing in plants using a CRISPR-Cas system. *Mol Plant*, **6**, 1975-1983.

**Xu, B., Ohtani, M., Yamaguchi, M., Toyooka, K., Wakazaki, M., Sato, M., Kubo, M., Nakano, Y., Sano, R., Hiwatashi, Y., Murata, T., Kurata, T., Yoneda, A., Kato, K., Hasebe, M. and Demura, T.** (2014) Contribution of NAC transcription factors to plant adaptation to land. *Science*, **343**, 1505-1508.

**Yaari, R., Noy-Malka, C., Wiedemann, G., Auerbach Gershovitz, N., Reski, R., Katz, A. and Ohad, N.** (2015) DNA METHYLTRANSFERASE 1 is involved in (m)CG and (m)CCG DNA methylation and is essential for sporophyte development in Physcomitrella patens. *Plant molecular biology*, **88**, 387-400.

**Zacharias, D.A. and Tsien, R.Y.** (2006) Molecular biology and mutation of green fluorescent protein. *Methods Biochem Anal*, **47**, 83-120.

**Zimmer, A.D., Lang, D., Buchta, K., Rombauts, S., Nishiyama, T., Hasebe, M., Van de Peer, Y., Rensing, S.A. and Reski, R.** (2013) Reannotation and extended community resources for the

genome of the non-seed plant Physcomitrella patens provide insights into the evolution of plant gene structures and functions. *BMC Genomics*, **14**, 498.

# 9 Danksagung

Mein besonderer Dank gilt Prof. Dr. Stefan Rensing für die Geduld und Unterstützung bei dieser Arbeit und die großartige Zeit, die ich in seiner Arbeitsgruppe verbringen durfte. Zudem danke ich ihm für die Erstkorrektur dieser Arbeit.

Prof. Dr. Alfred Batschauer danke ich für die Zweitkorrektur dieser Arbeit.

Dr. Katrin Heer und Prof. Dr. Annette Becker danke ich für ihren Beitrag zur Prüfungskommission.

Besonders danken möchte Dr. Kristian Ullrich, Dr. Mareike Schallenberg-Rüdinger, Dr. Christopher Grosche und Dr. Pierre-François Perroud, die immer Zeit für mich hatten und mich hervorragend unterstützt und begleitet haben. Dr. Pierre-François Perroud möchte ich außerdem für das Korrekturlesen dieser Arbeit danken.

Mein Dank gilt Marco Göttig und Rabea Meyberg, die mich bei vielen Experimenten unterstützt haben und meistens eine Lösung für die vielen kleinen und großen Herausforderungen gefunden haben.

Für die wichtigen Raucherpausen und die daraus entstandenen Diskussionen und Ideen möchte ich Dr. Katia Symeonidi, Lucas Schneider und Sebastian Hanke danken.

Ich danke allen Mitarbeitern der AG Rensing für die Unterstützung und die wunderbare Atmosphäre bei der Arbeit und auch außerhalb der Arbeit.

Ich danke meiner Frau, die mich trotz meiner manchmal schlechten Laune immer unterstützt hat und mir den Halt gegeben hat um diese Arbeit fertig zu stellen.

Mein Dank gilt vor allem auch meinen Eltern, die mir das Studium ermöglicht haben und die immer für mich da sind.

# 10 Curriculum vitae

Die Seite 87 (Lebenslauf) enthält persönliche Daten. Sie ist deshalb nicht Bestandteil der Online-Veröffentlichung.

# 11 Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quellen gekennzeichnet. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- beziehungsweise Beratungsdiensten (Promotionsberater oder anderer Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Die Bestimmungen der Promotionsordnung der Fakultät für Biologie der Universität Marburg sind mir bekannt, insbesondere weiß ich, dass ich vor Vollzug der Promotion zur Führung des Doktortitels nicht berechtigt bin.

Marburg (Lahn), den 15.03.2019

_____

Manuel Hiß